PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G01N 33/68 — A2

(11) International Publication Number: **WO 98/53323**

(43) International Publication Date: 26 November 1998 (26.11.98)

(21) International Application Number: PCT/GB98/01486

(22) International Filing Date: 22 May 1998 (22.05.98)

(30) Priority Data:
9710582.9 22 May 1997 (22.05.97) GB

(71) Applicant *(for all designated States except US)*: OXFORD GLYCOSCIENCES (UK) LTD. [GB/GB]; 10 The Quadrant, Barton lane, Abingdon Science Park, Abingdon OX14 3YS (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: PAREKH, Raj, Bhikhu [GB/GB]; Alchester House, Langford Lane, Nr Wendlebury, Oxfordshire OX6 0NS (GB). PRIME, Sally, Barbara [GB/GB]; Sunnybrook, 37 North Hinksey Village, Oxford OX2 0NA (GB). WEDD, Nick, Sinclair [GB/GB]; Sunnybrook, 37 North Hinksey Village, Oxford OX2 0NA (GB). TOWNSEND, Robert, Reid [US/GB]; 33 Norreys Avenue, Oxford OX1 4ST (GB).

(74) Agent: GILL JENNINGS & EVERY; Broadgate House, 7 Eldon Street, London EC2M 7LH (GB).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: A METHOD FOR DE NOVO PEPTIDE SEQUENCE DETERMINATION

(57) **Abstract**

A method for determining the amino acid sequence of an unknown peptide comprising (a) determining a molecular mass and an experimental fragmentation spectrum for the unknown peptide; (b) comparing the experimental fragmentation spectrum of the unknown peptide to theoretical fragmentation spectra calculated for a peptide library composed of all possible linear sequences of amino acids having a total mass that corresponds to the molecular mass of the unknown peptide; and (c) identifying a peptide in the peptide library having a theoretical fragmentation spectrum that matches most closely the fragmentation spectrum of the unknown peptide.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

A METHOD FOR DE NOVO PEPTIDE SEQUENCE DETERMINATION

FIELD OF THE INVENTION

The invention relates to a method for the determination of the precise linear sequence of amino acids in a peptide, polypeptide or protein, without recourse or reference to either a known pre-defined data base or to sequential amino acid residue analysis. As such, the method of the invention is a true, *de novo* peptide sequence determination method.

BACKGROUND OF THE INVENTION

The composition of a peptide (which term includes also polypeptide or protein) as a sequence of amino acids is well understood. Each peptide is uniquely defined by a precise linear sequence of amino acids. Knowledge of the precise linear arrangement or sequence of amino acids in a peptide is required for various purposes, including DNA cloning in which the sequence of amino acids provides information required for oligonucleotide probes and polymerase chain reaction ("PCR") primers. Knowledge of the exact sequence also allows the synthesis of peptides for antibody production, provides identification of peptides, aids in the characterization of recombinant products, and is useful in the study of post-translational modifications.

A variety of sequencing methods are available to obtain the amino acid sequence information. For example, a series of chemical reactions, e.g., Edman reactions, or enzymatic reactions, e.g., exo-peptidase reactions, are used to prepare sequential fragments of the unknown peptide. Either an analysis of the sequential fragments or a sequential analysis of the removed amino acids is used to determine the linear amino acid sequence of the unknown peptide. Typically, the Edman degradation chemistry is used in modern automated protein sequencers.

In the Edman degradation, a peptide is sequenced by degradation from the N-terminus using the Edman reagent, phenylisothiocyanate (PITC). The degradation process involves three steps, i.e., coupling, cleavage, and conversion. In the coupling step, PITC modifies the N-terminal residue of the peptide, polypeptide, or protein. An acid cleavage then cleaves the N-terminal amino acid in the form of an unstable anilinothiazolinone (ATZ) derivative, and leaves the peptide with a reactive N-terminus

and shortened by one amino acid. The ATZ derivative is converted to a stable phenylthiohydantoin in the conversion step for identification, typically with reverse phase high performance liquid chromatography (RP-HPLC). The shortened peptide is left with a free N-terminus that can undergo another cycle of the degradation reaction. Repetition of the cycle results in the sequential identification of each amino acid in the peptide. Because of the sequential nature of amino acid release, only one molecular substance can be sequenced at a time. Therefore, peptide samples must be extremely pure for accurate and efficient sequencing. Typically, samples must be purified with HPLC or SDS-PAGE techniques.

Although many peptide sequences have been determined by Edman degradation, currently, most peptide sequences are deduced from DNA sequences determined from the corresponding gene or cDNA. However, the determination of a protein sequence using a DNA sequencing technique requires knowledge of the specific nucleotide sequence used to synthesize the protein. DNA sequencing cannot be used where the nature of the protein or the specific DNA sequence used to synthesize the protein is unknown.

A peptide sequence may also be determined from experimental fragmentation spectra of the unknown peptide, typically obtained using activation or collision-induced fragmentation in a mass spectrometer. Tandem mass spectrometry (MS/MS) techniques have been particularly useful. In MS/MS, a peptide is first purified, and then injected into a first mass spectrometer. This first mass spectrometer serves as a selection device, and selects a target peptide of a particular molecular mass from a mixture of peptides, and eliminates most contaminants from the analysis. The target molecule is then activated or fragmented to form a mixture from the target or parent peptide of various peptides of a lower mass that are fragments of the parent. The mixture is then selected through a second mass spectrometer (i.e. step), generating a fragment spectrum.

Typically, in the past, the analysis of fragmentation spectra to determine peptide sequences has involved hypothesizing one or more amino acid sequences based on the fragmentation spectrum. In certain favorable cases, an expert researcher can interpret the fragmentation spectra to determine the linear amino acid sequence of an unknown peptide. The candidate sequences may then be compared with known amino acid sequences in protein sequence libraries.

In one strategy, the mass of each amino acid is subtracted from the molecular mass of the parent peptide to determine the possible molecular mass of a fragment, assuming that each amino acid is in a terminal position. The experimental fragment spectrum is then examined to determine if a fragment with such a mass is present. A score is generated for each amino acid, and the scores are sorted to generate a list of partial sequences for the next subtraction cycle. The subtraction cycle is repeated until subtraction of the mass of an amino acid leaves a difference of between -0.5 and 0.5, resulting in one or more candidate amino acid sequences. The highest scoring candidate sequences are then compared to sequences in a library of known protein sequences in an attempt to identify a protein having a sub-sequence similar or identical to the candidate sequence that generated the fragment spectrum.

Although useful in certain contexts, there are difficulties related to hypothesizing candidate amino acid sequences based on fragmentation spectra. The interpretation of fragmentation spectra is time-consuming, can generally be performed only in a few laboratories that have extensive experience with mass spectrometry, and is highly technical and often inaccurate. Human interpretation is relatively slow, and may be highly subjective. Moreover, methods based on peptide mass mapping are limited to peptide masses derived from an intact homogeneous peptide generated by specific, known proteolytic cleavage, and, thus, are not applicable in general to a mixture of peptides.

U.S. Patent No. 5,538,897 to Yates, III et al. provides a method of correlating the fragmentation spectrum of an unknown peptide with theoretical spectra calculated from described peptide sequences stored in a database to match the amino acid sequence of the unknown peptide to that of a described peptide. Known amino acid sequences, e.g., in a protein sequence library, are used to calculate or predict one or more candidate fragment spectra. The predicted fragment spectra are then compared with the experimentally-obtained fragment spectrum of the unknown protein to determine the best match or matches. Preferably, the mass of the unknown peptide is known. Sub-sequences of the various sequences in the protein sequence library are analyzed to identify those sub-sequences corresponding to a peptide having a mass equal to or within a given tolerance of the mass of the parent peptide in the fragmentation spectrum. For each sub-sequence having the proper mass, a predicted fragment spectrum can be calculated by calculating masses of various amino acid subsets of the candidate peptide. As a result,

a plurality of candidate peptides, each having a predicted fragment spectrum, is obtained. The predicted fragment spectra are then compared with the fragment spectrum obtained experimentally for the unknown protein, to identify one or more proteins having sub-sequences that are likely to be identical to the sequence of peptides that resulted in the experimentally-derived fragment spectrum. However, this technique cannot be used to derive the sequence of unknown, novel proteins or peptides having no sequence or sub-sequence identity with those pre-described or contained in such databases, and, thus, is not a *de novo* sequencing method.

Therefore, there remains a need for a true *de novo* sequencing method of determining the amino acid sequence of a peptide using mass spectrometry.

SUMMARY OF THE INVENTION

The present invention is directed to a method for generating a library of peptides, wherein each peptide in the library has a molecular mass corresponding to the same predetermined molecular mass. Typically, the library of peptides is then used to determine the amino acid sequence of an unknown peptide having the predetermined molecular mass. Preferably, the predetermined molecular mass used to generate the library is the molecular mass of the unknown peptide. Most preferably, the molecular mass of the unknown peptide is determined prior to the generation of the library using a mass spectrometer, such as a time-of-flight mass spectrometer.

The library is synthetic, i.e., not pre-described, and is typically generated each time a peptide is analyzed, based on the predetermined molecular mass of the unknown peptide. The library is generated by defining a set of all allowed combinations of amino acids that can be present in the unknown peptide, where the molecular mass of each combination corresponds to the predetermined molecular mass within the experimental accuracy of the device used to determine the molecular mass, allowing for water lost in peptide bond formation and for protonation, and generating an allowed library of all possible permutations of the linear sequence of amino acids in each combination in the set.

Generally, the present invention is directed to a method for determining the amino acid sequence of an unknown peptide, which comprises determining a molecular mass and an experimental fragmentation spectrum for the unknown peptide, comparing the

experimental fragmentation spectrum of the unknown peptide to theoretical fragmentation spectra calculated for each individual member of an allowed synthetic peptide library, where the allowed peptide library is of the type described above, and identifying a peptide in the peptide library having a theoretical fragmentation spectrum that matches most closely the fragmentation spectrum of the unknown peptide, from which it is inferred that the amino acid sequence of the identified peptide in the allowed library represents the amino acid sequence of the unknown peptide.

The molecular mass for the unknown peptide may be determined by any means known in the art, but is preferably determined with a mass spectrometer. Allowed combinations of amino acids are chosen from a set of allowed amino acids that typically comprises the natural amino acids, i.e., tryptophan, arginine, histidine, glutamic acid, glutamine, aspartic acid, leucine, threonine, proline, alanine, tyrosine, phenylalanine, methionine, lysine, asparagine, isoleucine, cysteine, valine, serine, and glycine, but may also include other amino acids, including, but not limited to, non-natural amino acids and chemically modified derivatives of the natural amino acids, e.g., carbamidocysteine and deoxymethionine. Allowed combinations of amino acids are then calculated using one or more individual members of this set of amino acids, allowing for known mass changes associated with peptide bond formation, such that the total mass of each allowed combination corresponds to the predetermined mass of the unknown peptide to within the experimental accuracy to which this molecular mass of the unknown peptide was calculated, typically about 30 ppm. The set of allowed combinations is most easily calculated using an appropriately programmed computer. The allowed peptide library is assembled by permutation in all possible linear combinations of each allowed amino acid composition, and is also most easily constructed using an appropriately programmed computer. It should be noted that the term "allowed" with respect to amino acid combinations and libraries of peptides refers to combinations and libraries specific to the unknown peptide under investigation. The peptide library is constructed from the amino acid combinations, which in turn are calculated from the experimentally determined molecular mass. As unknown peptides of different mass are investigated, so different combinations of amino acids are allowed, and hence each unknown peptide of unique molecular mass gives rise to a unique peptide library.

The present invention constrains the allowed library, i.e. limits the number of possible sequences. In the broadest aspect of the invention, this constraint is achieved by determining a molecular mass for the peptide whose sequences is to be determined, i.e. the unknown peptide.

According to preferred embodiments of the invention, information, e.g. available from the experimental fragmentation spectrum of the unknown peptide, can be used to put further constraints on the number of possible sequences of amino acids in the peptide library. For example, the immonium ion region of the mass spectrum used to determine the molecular mass may also be used to identify amino acids contained in the unknown peptide. Alternatively or in addition, the two N-terminal amino acids may be identified from the b_2/a_2 ion pairs. For example, the two N-terminal amino acids may be deduced from the prominent signals of the b_2 and a_2 ions. In particular, the identity of the signals may be determined by recognition of signals separated by 27.98 a.m.u. (corresponding to CO) in the region of the spectrum which includes the mass of all possible combinations of modified and unmodified amino acids. Further, based on the use of enzyme treatment, e.g. with a protease such as papain, chymotrypsin or trypsin, the C-terminal residue of any peptide in the spectrum is determined as either arginine or lysine, and this may be confirmed or identified from the recognition of signals at 175.11 and 147.11 respectively. Alternatively, C-terminals containing basic amino acids can be identified by recognition of the predicted y_1 ion. The spectrum can be interpreted to identify the next amino acids.

Another means of applying a constraint on the allowed library of amino acids is to obtain partial internal sequence information, e.g. by identifying the y series of ions with appropriate defined accuracy of mass measurement. In particular, a computer programme may be used to recognise at least three sequential signals separated by the mass of all possible modified and unmodified amino acid residues. The differences between these signals allows identification of a sequence of two amino acids. Most preferably, the molecular mass of the unknown peptide and at least one other experimental parameter, e.g. as given above, are used as constraints in initially generating the library of allowed peptides.

The nature of the fragmentation process from which the theoretical fragmentation spectrum is calculated for every peptide in the allowed library may be of any type known in the art, such as a mass spectrum or a protease or chemical fragmentation spectrum.

Preferably, both the molecular mass and the fragmentation spectrum for the unknown peptide are obtained from a tandem mass spectrometer. The amino acid sequence of the peptide from the allowed library of peptides, having a calculated fragmentation spectrum that best fits the experimental fragmentation spectrum of the unknown peptide, corresponds to the amino acid sequence of the unknown peptide.

Although not required, the experimental fragmentation spectrum is generally normalized. A factor that is an indication of closeness-of-fit between the experimental fragmentation spectrum of the unknown peptide and each of the theoretical fragmentation spectra calculated for the peptide library may then be calculated to determine which of the theoretical fragmentation spectra best fits the experimental fragmentation spectrum. Preferably, peak values in the fragmentation spectra having an intensity greater than a predetermined threshold value are selected when calculating the indication of closeness-of-fit. The theoretical fragmentation spectrum that best fits the experimental fragmentation spectrum corresponds to the amino acid sequence in the allowed library that matches that of the unknown peptide.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow chart of the method of the invention.

Figs. 2a and 2b are flow charts of alternative preferred embodiments of the invention.

Fig. 3 is the experimental mass spectrum used to determine the molecular mass of unknown Peptide X.

Fig. 4 is the immonium ion region of the mass spectrum shown in Fig. 3, and identifies amino acids contained in unknown Peptide X.

Fig. 5 is the experimental fragmentation mass spectrum of Peptide X.

Fig. 6 is the experimental mass spectrum used to determine the molecular mass of a Peptide Y.

Fig. 7 is the immonium ion region of the mass spectrum shown in Fig. 6, and identifies amino acids contained in Peptide Y.

Fig. 8 is the experimental tandem mass spectrum of Peptide Y.

Fig. 9 is the experimental mass spectrum used to determine the molecular mass of a Peptide Z.

Fig. 10 is the immonium ion region of the mass spectrum shown in Fig. 9, and identifies amino acids contained in Peptide Z.

Fig. 11 is the experimental tandem mass spectrum of Peptide Z.

DETAILED DESCRIPTION OF THE INVENTION

The present invention is directed to a *de novo* method for determining the sequence of an unknown peptide without reference to any experimentally determined peptide or nucleotide sequence, and without recourse to a sequential and step-wise identification and ordering of individual amino acid residues, such as the Edman degradation process or interpretation of conventional mass spectrometry fragmentation patterns. In the method of the invention, a library of theoretical peptide sequences is generated from a predetermined molecular mass, preferably the experimentally determined molecular mass of an unknown peptide. As such, this library must contain the amino acid sequence of the unknown peptide, as well as that of any other peptide having the predetermined molecular mass. The precise amino acid sequence of the unknown is identified by applying standard correlation functions to select that peptide from the synthetic library whose calculated, i.e., theoretical, fragmentation spectrum most closely matches the fragmentation pattern of the unknown. In the preferred embodiment, the fragmentation spectrum is a mass spectrum and the correlation method is the function described in U.S. Patent No. 5,538,897, the contents of which are incorporated herein in their entirety by reference. Preferably, the theoretical fragmentation spectra are generated and matched to the fragmentation pattern of the unknown using an appropriately programmed computer.

The invention may be better understood by reference to the flow chart provided in Fig. 1. Where the peptide is a protein or large polypeptide, the protein or large polypeptide may be cleaved to form a peptide pool by means well known in the art. The unknown peptide ("Peptide X") is then separated from the pool by HPLC or any other means known in the art, preferably mass spectrometry, and the molecular mass of Peptide X is determined. Although there are a number of methods for determining the molecular mass of Peptide X, the preferred method is again mass spectrometry.

A set of amino acids that theory or experimental results teach may be included in Peptide X is then defined for consideration in determining the sequence of Peptide X.

The defined set of amino acids may include modified or unnatural amino acids in addition to natural amino acids.

Typically, the method of the invention requires a "naked" peptide when determining the amino acid sequence. Therefore, the peptide should be free of any individual amino acids that are covalently modified by post-translational modification, such as, e.g., glycosylation, which involves the attachment of carbohydrate to the side chain of certain amino acids. Where the method of the invention is used to determine the amino acid sequence of a post-translationally modified peptide, the modifications are typically removed from the peptide prior to the analysis, taking due care to leave the peptide intact. Methods for removing post-translational modifications from peptides are well known in the art, and include, for example, the removal of N-linked carbohydrates with enzymes, such as peptide-N-glycosidase F (PNGase F), endo-glycosidases, mixtures of exo-glycosidases, etc., and the removal of phosphate modification with phosphatases. In addition, other techniques for removing modifications occasionally found on peptides are well known in the art. However, where a specific modification to a specific amino acid is known to be present in the unknown peptide, the modified amino acid may be included in the defined set of amino acids that theory or experimental results teach may be included in Peptide X, and, thus, the sequence of the peptide containing the modified peptide may be determined with the method of the present invention.

All combinations of amino acids having a total mass equal to the measured mass of Peptide X are calculated, allowing for water lost in determining peptide links, protonation, etc. Any individual amino acid may be included as part of any given combination at any integral stoichiometry up to the amount consistent with the mass determined for Peptide X. These combinations comprise all of the allowed combinations of amino acids combinations for Peptide X, and, therefore, the actual amino acid compositions of Peptide X will be represented in one and only one of these combinations. Furthermore, these combinations are generally peptide-specific.

An allowed library of linear peptides is then constructed from the allowed combinations of amino acids. The allowed library is constructed by generating all possible linear permutations of the sequence of amino acids in each combination, using all the amino acids in each combination. The allowed library comprises all such permutations of the amino acids, and therefore must include Peptide X. The allowed library of peptides

having the same molecular mass as Peptide X is typically constructed independently and *ab initio* for each new unknown peptide that is sequenced. That is, a new library is typically constructed as part of each analysis, and for only that analysis. However, as will be clear to one of ordinary skill in the art, once a library of all peptides having a given molecular mass has been constructed, that library may be used for the determination of the amino acid sequence of any other peptide of that particular molecular mass.

This differs fundamentally from existing data base approaches in which a single data base of known sequences, which is subject to periodic updates and refinements based on the availability of experimentally determined sequences, is used for all analyses. As a result, with the method of the present invention, the determination of new and previously unknown peptides sequences that are not present in any experimentally determined peptide sequence library is possible by direct peptide analysis in a non-step-wise, operator-independent automated process. In addition, the method of the invention is not constrained to the conventional twenty amino acids, or to their conventional modifications.

In a preferred embodiment, as shown in the flow chart provided in Fig. 2a, additional information relating to Peptide X is used to place constraints on the allowed combinations of amino acids and/or allowed peptide sequences in the library, and, thus, reduce the number of possible sequences. Useful information related to Peptide X includes, but is not limited to, partial amino acid composition. For example, the mass spectrum used to determine the mass of Peptide X may include fragments that can be used to identify specific amino acids present in Peptide X. Where it is known that certain amino acids are definitely present in Peptide X, constraints are placed on the allowed combinations and allowed library by requiring the identified amino acids to be present in all combinations and, thus, in every peptide present in the library.

Fig. 2b illustrates a system whereby more than one constraint is put on the library of possible linear sequences. By way of illustration only, for each peptide to be analysed (whether purified or present in a mixture), information on its mass (e.g. by MALDI-MS) and a tandem mass spectrum from it (e.g. by FSI-tandem MS) are obtained. The tandem mass spectrum can then be interpreted in an automated manner, to obtain certain information about the unknown peptide. Suitable software evaluates the following information, when possible, from a tandem MS spectrum in an automated manner:

i. Information on amino acids contained in the peptide by analysing the immonium ions region.

ii. Identification of the two N-terminal amino acids by identifying the b_2/a_2 ion pairs.

iii. Based on the use of trypsin, the C-terminal residue must be lysine or arginine. These can be identified in the spectrum and the spectrum interpreted to find the next amino acids.

iv. Partial internal sequence information can be obtained by identifying y series of ions with defined accuracy of mass measurement at ≤ 100 ppm.

A discussion of manual spectrum interpretation is provided in Medzihradsky and Burlingame, A Companion to Methods in Enzymology 6: 284-303 (1994).

Again with reference to Figs. 1 and 2, the allowed library, which has preferably been constrained, is then used as the basis for generating theoretical fragmentation patterns that are compared to the experimental fragmentation pattern obtained for Peptide X. The fragmentation patterns may be obtained by any suitable means known in the art. Preferably, the fragmentation patterns are mass spectra, and the method used to match the theoretical and experimental mass spectra is that disclosed in U.S. Patent No. 5,538,897. However, protease or chemical fragmentation, coupled to HPLC separation of the fragments, may also be used to obtain the experimental fragmentation patterns.

Preferably, in a determination of the amino acid sequence of Peptide X, the molecular mass of Peptide X is determined with high accuracy, typically, to within about 30 ppm (parts per million). An example of such a spectrum is provided in Fig. 3, where the molecular mass of Peptide X is determined from the peak at 774.3928 daltons. In addition, as a result of the partial fragmentation of Peptide X that can occur, fragments that identify certain amino acids that are contained in Peptide X are also observed, allowing the peptide library to be constrained. An example of this portion of the mass spectrum for Peptide X is provided in Fig. 4.

Peptide X is then subjected to collision-induced dissociation in a mass spectrometer. The parent peptide and its fragments are then introduced into the second mass spectrometer that provides an intensity or count and the mass to charge ratio, m/z, for each of the fragments in the fragment mixture. Each fragment ion is represented in a bar graph in which the abscissa value is m/z and the ordinate value is the intensity. A

variety of mass spectrometer types can be used, including, but not limited to, triple-quadrapole mass spectrometry, Fourier-transform cyclotron resonance mass spectrometry, tandem time-of-flight mass spectrometry, and quadrapole ion trap mass spectrometry.

The experimental fragment spectrum is then compared to the mass spectra predicted for the sequences of the allowed library, to identify one or more predicted mass spectra that closely match the experimental mass spectrum. Because the allowed library includes all permutations of amino acid sequences that have a total mass corresponding to that of Peptide X, Peptide X must be represented in the allowed library.

The predicted fragmentation spectra may be obtained and compared to the experimental fragmentation spectrum by employing a method that involves first normalizing the experimental fragmentation spectrum. This may be accomplished by converting the experimental fragmentation spectrum to a list of masses and intensities. The peak values for Peptide X are removed, and the square root of the remaining intensity values is calculated, and normalized to a maximum value of 100. The 200 most intense ions are divided into ten mass regions, and the maximum intensity within each region is again normalized to 100. Each ion within 3.0 daltons of its neighbour on either side is given an intensity value equal to the greater of the intensity of the ion or that of its neighbour. Other normalization methods can be used, and it is possible to perform the analysis without normalizing. However, in general, normalization is preferred. In particular, maximum normalized values, the number of intense ions, the number of mass regions, and the size of the window for assuming the intensity value of a near neighbour may all be independently varied to larger or smaller values.

A fragment mass spectrum is predicted for each of the candidate sequences. The fragment mass spectrum is predicted by calculating the fragment ion masses for the type b and y ions for the amino acid sequence. When a peptide is fragmented and the charge is retained on the N-terminal cleavage fragment, the resulting ion is labelled as a b-type ion. If the charge is retained on the c-type terminal fragment, it is labelled a y-type ion. Masses for type b ions were calculated by summing the amino acid masses and adding the mass of a proton. Masses for type y ions were calculated by summing, from the c-terminus, the masses of the amino acids and adding the mass of water and a proton to

the initial amino acid. In this way, it is possible to calculate an m/z value for each fragment.

However, in order to provide a predicted mass spectrum, it is also necessary to assign an intensity value for each fragment. Although it is often possible to predict, on a theoretical basis, an intensity value for each fragment, this procedure is difficult, and it has been found useful to assign intensities in the following fashion. The value of 50.0 is assigned to each b and y ion. To masses of 1 dalton on either side of the fragment ion, an intensity of 25.0 is assigned. Peak intensities of 10.0 are assigned at mass peaks 17.0 and 18.0 daltons below the m/z of each b and y ion location, to account for both NH_3 and H_2O loss, and peak intensities of 10.0 are assigned to mass peaks 28.0 daltons below each type b ion location, to account for CO loss.

After calculation of predicted m/z values and assignment of intensities, it is preferred to calculate a measure of closeness-of-fit between the predicted mass spectra and the experimentally-derived fragment spectrum. A number of methods for calculating closeness-of-fit are available. For example, a two-step method may be used that includes calculating a preliminary closeness-of-fit score, referred to here as S_p, and calculating a correlation function for the highest-scoring amino acid sequences. In the preferred embodiment, S_p is calculated using the following formula:

$$S_p = (\Sigma i_m) * n_i * (1+\beta)*(1-\rho)/n_t \qquad (1)$$

where i_m are the matched intensities, n_i are the number of matched fragment ions, β is the type b and y ion continuity, ρ is the presence of immonium ions and their respective amino acids in the predicted sequence, and n_t is the total number of fragment ions. The factor β evaluates the continuity of a fragment ion series. If there is a fragment ion match for the ion immediately preceding the current type b or y ion, β is incremented by 0.075 from an initial value of 0.0. This increases the preliminary score for those peptides matching a successive series of type b and y ions, since extended series of ions of the same type are often observed in MS/MS spectra. The factor p evaluates the presence of immonium ions in the low mass end of the mass spectrum.

The detection of immonium ions may be used diagnostically to determine the presence of certain types of amino acids in the sequence. For example, if immonium ions are present at 110.0, 120.0, or 136.0 + 1.0 daltons in the processed data file of the unknown peptide with normalized intensities greater than 40.0, indicating the presence of histidine, phenylalanine, and tyrosine respectively, then the sequence under evaluation is checked for the presence of the amino acid indicated by the immonium ion. The preliminary score, S_p, for the peptide is either increased or decreased by a factor of 1-ρ, where ρ is the sum of the penalties for each of the three amino acids whose presence is indicated in the low mass region. Each individual ρ can take on the value of -0.15 if there is a corresponding low mass peak, and the amino acid is not present in the sequence, +0.15 if there is a corresponding low mass peak and the amino acid is present in the sequence, or 0.0 if the low mass peak is not present. The total penalty can range from -0.45, where all three low mass peaks are present in the spectrum, but are not present in the sequence, to +0.45, where all three low mass peaks are present in the spectrum, and are present in the sequence.

Following the calculation of the preliminary closeness-of-fit score, S_p, the predicted mass spectra having the highest S_p scores are selected for further analysis using the correlation function. The number of candidate predicted mass spectra that are selected for further analysis will depend largely on the computational resources and time available.

For purposes of calculating the correlation function, the experimentally-derived fragment spectrum is typically preprocessed in a fashion somewhat different from preprocessing employed before calculating S_p. For purposes of the correlation function, the precursor ion is removed from the spectrum, and the spectrum is divided into 10 sections. Ions in each section are then normalized to 50.0. The section-wise normalized spectra are then used for calculating the correlation function. The discrete correlation between the two functions may be calculated as:

$$R_\tau = \sum_{i=0}^{n-1} x_i y_i + \tau \qquad (2)$$

where τ is a lag value. The discrete correlation theorem states that the discrete correlation of two real functions x and y is one member of the discrete Fourier transform pair

$$R_\tau = X_\tau Y^*\tau \qquad (3)$$

where X(t) and Y(t) are the discrete Fourier transforms of x(i) and y(i), and the Y^* denotes complex conjugation. Therefore, the cross-correlations can be computed by Fourier transformation of the two data sets using the fast Fourier transform (FFT) algorithm, multiplication of one transform by the complex conjugate of the other, and inverse transformation of the resulting product.

The predicted spectra as well as the pre-processed unknown spectrum may be zero-padded to 4096 data points, since the MS/MS spectra are not periodic, as intended by the correlation theorem, and the FFT algorithm requires N to be a integer power of two, so the resulting end effects need to be considered. The final score attributed to each candidate peptide sequence is R(0) minus the mean of the cross-correlation function over the range $-75<t<75$. This modified "correlation parameter", described in Powell and Heiftje, Anal. Chim. Acta, 100:313-327 (1978), shows better discrimination over just the spectral correlation coefficient R(0). The raw scores are normalized to 1.0. Preferably, the output includes the normalized raw score, the candidate peptide mass, the unnormalized correlation coefficient, the preliminary score, the fragment ion continuity β, the immonium ion factor τ, the number of type b and y ions matched out of the total number of fragment ions, their matched intensities, the protein accession number, and the candidate peptide sequence.

The correlation function can be used to select automatically one of the predicted mass spectra as corresponding to the experimentally-derived fragment spectrum. Preferably, however, a number of sequences from the library are output and final selection of a single sequence is done by a skilled operator.

Depending on the computing and time resources available, it may be advantageous to employ data-reduction techniques. Preferably, these techniques will emphasize the most informative ions in the spectrum while not unduly affecting search speed. One technique involves considering only some of the fragment ions in the MS/MS spectrum, which, for a peptide, may contain as many as 3,000 fragment ions. According

to one data reduction strategy, the ions are ranked by intensity, and some fraction of the most intense ions is used for comparison. Another approach involves subdividing the spectrum into a small number of regions, e.g., about 5, and using the 50 most intense ions in each region as part of the data set. Yet another approach involves selecting ions based on the probability of those ions being sequence ions. For example, ions could be selected which exist in mass windows of 57 through 186 daltons, i.e., the range of mass increments for the 20 common amino acids from glycine to tryptophan that contain diagnostic features of type b or y ions, such as losses of 17 or 18 daltons, corresponding to ammonia and water, or a loss of 28 daltons, corresponding to CO .

A number of different scoring algorithms can be used for determining preliminary closeness-of-fit or correlation. In addition to scoring based on the number of matched ions multiplied by the sum of the intensity, scoring can be based on the percentage of continuous sequence coverage represented by the sequence ions in the spectrum. For example, a 10 residue peptide will potentially contain 9 each of b and y type sequence ions. If a set of ions extends from B_1 to B_9, then a score of 100 is awarded, but if a discontinuity is observed in the middle of the sequence, such as a missing B_5 ion, a penalty is assessed. The maximum score is awarded for an amino acid sequence that contains a continuous ion series in both the b and y directions.

In the event that the described scoring procedures do not delineate an answer, an additional technique for spectral comparison can be used in which the database is initially searched with a molecular weight value and a reduced set of fragment ions. Initial filtering of the database occurs by matching sequence ions, and generating a score with one of the methods described above. The resulting set of answers will then undergo a more rigorous inspection process using a modified full MS/MS spectrum.

For the second stage analysis, one of several spectral matching approaches developed for spectral library searching is used. This will require generating a "library spectrum" for the peptide sequence, based on the sequence ions predicted for that amino acid sequence. Intensity values for sequence ions of the "library spectrum" will be obtained from the experimental spectrum. If a fragment ion is predicted at m/z 256, then the intensity value for the ion in the experimental spectrum at m/z 256 will be used as the intensity of the ion in the predicted spectrum. Thus, if the predicted spectrum is identical to the "unknown" spectrum, it will represent an ideal spectrum. The spectra will then be

compared using a correlation function. In general, it is believed that the majority of computational time for the above procedure is spent in the iterative search process. By multiplexing the analysis of multiple MS/MS spectra in one pass through the database, an overall improvement in efficiency will be realized. In addition, the mass tolerance used in the initial pre-filtering can affect search times by increasing or decreasing the number of sequences to analyze in subsequent steps.

Another approach to speeding up searches involves a binary encryption scheme where the mass spectrum is encoded as peak/no peak at every mass depending on whether the peak is above a certain threshold value. If intensive use of a protein sequence library is contemplated, it may be possible to calculate and store predicted mass values of all sub-sequences within a predetermined range of masses so that at least some of the analysis can be performed by table look-up rather than calculation.

EXAMPLES

The following non-limiting examples are merely illustrative of the preferred embodiments of the present invention, and are not to be construed as limiting the invention, the scope of which is defined by the appended claims.

EXAMPLE 1.

The amino acid sequence of unknown Peptide X was determined using the method of the invention. The molecular mass of Peptide X was first determined using a matrix-assisted laser-description time-of-flight mass spectrometer (Voyager Elite, manufactured by Perseptive Biosystems) with delayed extraction and post source decay. As shown in Fig. 3, the mass of the protonated form of Peptide X form is 774.3928 daltons, which indicates a mass of 773.3928 daltons for Peptide X.

The set of amino acids that are possibly part of Peptide X were then defined for consideration in the analysis. The defined set of amino acids with the molecular mass of each amino acid less the mass of the one water molecule lost during peptide bond formation is provided below. The molecular masses are given in daltons or a.m.u.

	tryptophan	=	186.079313	carbamido cysteine	=	160.03065
	arginine	=	156.10111	phenylalanine	=	147.068414
	histidine	=	137.058912	methionine	=	131.04085
	glutamic acid	=	129.042593	lysine	=	128.094963
5	glutamine	=	128.058577	asparagine	=	114.042927
	aspartic acid	=	115.026943	isoleucine	=	113.084064
	leucine	=	113.084064	cysteine	=	103.009185
	threonine	=	101.047678	valine	=	99.068414
	proline	=	97.052764	serine	=	87.032028
10	alanine	=	71.037114	glycine	=	57.021464
	tyrosine	=	163.063328			

The allowed combinations of amino acids for Peptide X were determined by first determining the molecular mass of Peptide X, as described above, to an experimental 15 accuracy of 30 ppm (parts per million). Therefore, each allowed combination of amino acids in the allowed library must have a total mass of 773.3928 ± 30 ppm. In addition to providing the molecular mass of Peptide X, the first mass spectrum also confirmed the presence of certain amino acids in Peptide X. The immonium region of this mass spectrum, which shows the presence of these amino acids, is given in Fig. 4. In particular, 20 the immonium region of the spectrum indicates the presence of arginine with a characteristic mass of 174.988, leucine/isoleucine with a characteristic mass at 85.8851 (these amino acids have the same mass, and are therefore not distinguishable by mass alone), histidine with a characteristic mass at 109.823, and tyrosine with a characteristic mass at 135.915. Therefore, it was possible to constrain the allowed library to sets 25 containing arginine, leucine/isoleucine, histidine, and tyrosine, having a total molecular mass of 773.3928 ± 30 ppm.

To determine the sets of amino acids that have a total molecular mass of 773.3928 ± 30 ppm, the following equation was applied:

$$MM_x = \Sigma \text{ (histidine) + (tyrosine) + (leucine/isoleucine) +}$$
30 $$\text{(arginine) + } (H_2O) + (aa_1) + \text{----+ } (aa_n),$$

where aa_1 ---- aa_n are any of the allowed amino acids, other than arginine, isoleucine, histidine, and tyrosine. The only combinations of amino acids that can have a total molecular mass of 773.3928 + 30 ppm are as follows:

1) tryptophan, arginine, leucine/isoleucine, histidine, and tyrosine.

2) glutamic acid, glycine, arginine, leucine/isoleucine, histidine, and tyrosine.

3) alanine, aspartic acid, arginine, leucine/isoleucine, histidine, and tyrosine.

These combinations constitute the allowed sets of amino acids for Peptide X.

In addition, Peptide X was obtained by a tryptic cleavage, and, therefore, from the accepted specificity of trypsin, Peptide X must also have lysine or arginine as its carboxy terminal amino acid. With this constraint, the allowed library of linear peptides was constructed from all individual linear permutations of combinations 1, 2, and 3. The allowed library includes 528 linear peptides, one set of 264 peptides containing isoleucine (shown below) and a corresponding set of 264 peptides in which isoleucine is replaced by leucine (not shown).

1) YIHWR	54) HIYGER	107) HGYEIR	160) DYHIAR	213) IADYHR
2) IYHWR	55) YIGHER	108) GHYEIR	161) HDYIAR	214) AIDYHR
3) YHIWR	56) IYGHER	109) YEGHIR	162) DHYIAR	215) DAIYHR
4) HYIWR	57) YGIHER	110) EYGHIR	163) IHDYAR	216) ADIYHR
5) IHYWR	58) GYIHER	111) YGEHIR	164) HIDYAR	217) YHDAIR
6) HIYWR	59) IGYHER	112) GYEHIR	165) IDHYAR	218) HYDAIR
7) YIWHR	60) GIYHER	113) EGYHIR	166) DIHYAR	219) YDHAIR
8) IYWHR	61) YHGIER	114) GEYHIR	167) HDIYAR	220) DYHAIR
9) YWIHR	62) HYGIER	115) HEGYIR	168) DHIYAR	221) HDYAIR
10) WYIHR	63) YGHIER	116) EHGYIR	169) YIHADR	222) DHYAIR
11) IWYHR	64) GYHIER	117) HGEYIR	170) IYHADR	223) YHADIR
12) WIYHR	65) HGYIER	118) GHEYIR	171) YHIADR	224) HYADIR
13) YHWIR	66) GHYIER	119) EGHYIR	172) HYIADR	225) YAHDIR
14) HYWIR	67) IHGYER	120) GEHYIR	173) IHYADR	226) AYHDIR
15) YWHIR	68) HIGYER	121) IHEGYR	174) HIYADR	227) HAYDIR

16) WYHIR	69) IGHYER	122) HIEGYR	175) YIAHDR	228) AHYDIR
17) HWYIR	70) GIHYER	123) IEHGYR	176) IYAHDR	229) YDAHIR
18) WHYIR	71) HGIYER	124) EIHGYR	177) YAIHDR	230) DYAHIR
19) IHWYR	72) GHIYER	125) HEIGYR	178) AYIHDR	231) YADHIR
20) HIWYR	73) YIEGHR	126) EHIGYR	179) IAYHDR	232) AYDHIR
21) IWHYR	74) IYEGHR	127) IHGEYR	180) AIYHDR	233) DAYHIR
22) WIHYR	75) YEIGHR	128) HIGEYR	181) YHAIDR	234) ADYHIR
23) HWIYR	76) EYIGHR	129) IGHEYR	182) HYAIDR	235) HDAYIR
24) WHIYR	77) IEYGHR	130) GIHEYR	183) YAHIDR	236) DHAYIR
25) YIHEGR	78) EIYGHR	131) HGIEYR	184) AYHIDR	237) HADYIR
26) IYHEGR	79) YIGEHR	132) GHIEYR	185) HAYIDR	238) AHDYIR
27) YHIEGR	80) IYGEHR	133) IEGHYR	186) AHYIDR	239) DAHYIR
28) HYEGR	81) YGIEHR	134) EIGHYR	187) IHAYDR	240) ADHYIR
29) IHYEGR	82) GYIEHR	135) IGEHYR	188) HIAYDR	241) IHDAYR
30) HIYEGR	83) IGYEHR	136) GIEHYR	189) IAHYDR	242) HIDAYR
31) YIEHGR	84) GIYEHR	137) EGIHYR	190) AIHYDR	243) IDHAYR
32) IYEHGR	85) YEGIHR	138) GEIHYR	191) HAIYDR	244) DIHAYR
33) YEIHGR	86) EYGIHR	139) HEGIYR	192) AHIYDR	245) HDIAYR
34) EYIHGR	87) YGEIHR	140) EHGIYR	193) YIDAHR	246) DHIAYR
35) IEYHGR	88) GYEIHR	141) HGEIYR	194) IYDAHR	247) IHADYR
36) EIYHGR	89) EGYIHR	142) GHEIYR	195) YDIAHR	248) HIADYR
37) YHEIGR	90) GEYIHR	143) EGHIYR	196) DYIAHR	249) IAHDYR
38) HYEIGR	91) IEGYHR	144) GEHIYR	197) IDYAHR	250) AIHDYR
39) YEHIGR	92) EIGYHR	145) YIHDAR	198) DIYAHR	251) HAIDYR
40) EYHIGR	93) IGEYHR	146) IYHDAR	199) YIADHR	252) AHIDYR
41) HEYIGR	94) GIEYHR	147) YHIDAR	200) IYADHR	253) IDAHYR
42) EHYIGR	55) EGIYHR	148) HYIDAR	201) YAIDHR	254) DIAHYR
43) IHEYGR	96) GEIYHR	149) IHYDAR	202) AYIDHR	255) IADHYR
44) HIEYGR	97) YHEGIR	150) HIYDAR	203) IAYDHR	256) AIDHYR
45) IEHYGR	98) HYEGIR	151) YIDHAR	204) AIYDHR	257) DAIHYR
46) EIHYGR	99) YEHGIR	152) IYDHAR	205) YDAIHR	258) ADIHYR
47) HEIYGR	100) EYHGIR	153) YDIHAR	206) DYAIHR	259) HDAIYR

48) EHIYGR	101) HEYGIR	154) DYIHAR	207) YADIHR	260) HADIYR
49) YIHGER	102) EHYGIR	155) IDYHAR	208) AYDIHR	261) HADIYR
50) IYHGER	103) YHGEIR	156) DIYHAR	209) DAYIHR	262) AHDIYR
51) YHIGER	104) HYGEIR	157) YHDIAR	210) ADYIHR	263) DAHIYR
52) HYIGER	105) YGHEIR	158) HYDIAR	211) IDAYHR	264) ADHIYR
53) IHYGER	106) GYHEIR	159) YDHIAR	212) DIAYHR	

The method of U.S. Patent No. 5,538,897 was then used to match Peptide X to this library by MS/MS. The experimental tandem mass spectrum of Peptide X is shown in Fig. 5, and the 10 top ranking peptides matched to this spectrum are provided below. It was determined that the sequence of Peptide X is that of the top ranked peptide, AHYDIR.

Rank/Sp	(M+H)	Cn	C*10^4	Sp	Ions	Reference	Peptide
1/1	774.9	1.0000	1.8118	491.0	11/15	p(228)	(-)AHYDIR
2/3	774.9	0.9308	1.6864	386.2	10/15	p(238)	(-)AHDYIR
3/2	774.9	0.8012	1.4516	414.3	10/15	p(227)	(-)HAYDIR
4/5	774.9	0.7319	1.3262	320.5	9/15	p(237)	(-)HADYIR
5/1	774.9	0.7168	1.2987	491.0	11/15	p(186)	(-)AHYIDR
6/12	774.9	0.6131	1.1108	248.3	9/15	p(226)	(-)AYHDIR
7/3	774.9	0.6033	1.0930	386.2	10/15	p(192)	(-)AHIYDR
8/9	774.9	0.5878	1.0651	264.1	9/15	p(225)	(-)YAHDIR
9/50	774.9	0.5850	1.0599	156.5	7/15	p(219)	(-)YDHAIR
10/14	774.9	0.5825	1.0553	247.9	9/15	p(217)	(-)YHDAIR

EXAMPLE 2.

The amino acid sequence of Peptide Y, a known, standard peptide, was determined using the method of the invention, as applied to Peptide X in Example 1. Peptide Y has the following amino acid sequence: YGGFIRR. The molecular mass of Peptide Y was determined to be 868.4719 to an experimental accuracy of 30 ppm from the mass spectrum shown in Fig. 6. The masses at 1296.6854 and 1570.6774 are from internal standards, added to allow instrument calibration.

The set of amino acids that are possibly part of Peptide Y were then defined for consideration in the analysis. The defined set of amino acids with the molecular mass of each amino acid less the mass of the one water molecule lost during peptide bind formation is the same as those used in Example 1.

As the mass of Peptide Y was measured as 868.4719 to an experimental accuracy of ± 30 ppm, each allowed amino acid combination must therefore have a total mass equal to 868.4719 ± 30 ppm. In addition, from the immonium ion region of the PSD trace from Fig. 6, shown in Fig. 7, it was determined that Peptide Y must also contain the following amino acids: tyrosine with a characteristic mass at 136.027, phenylalanine with a characteristic mass at 120.071, arginine with a characteristic mass at 175.00, and leucine or isoleucine with a characteristic mass at 85.9225.

Application of the equation in Example 1 demonstrated that only the following combinations of amino acids are allowed for Peptide Y:

1) Tyrosine, phenylalanine, arginine, asparagine, and arginine.
2) Tyrosine, phenylalanine, arginine, arginine, leucine/isoleucine, glycine, and glycine.
3) Tyrosine, phenylalanine, arginine, leucine/isoleucine, alanine, alanine, and glutamine.
4) Tyrosine, phenylalanine, arginine, leucine/isoleucine, glycine, valine, and asparagine
5) Tyrosine, phenylalanine, arginine, leucine/isoleucine, glycine, glycine, glycine, and valine
6) Tyrosine, phenylalanine, arginine, leucine/isoleucine, glycine, alanine, alanine, and alanine.

These combinations constitute the allowed set of amino acid combinations for Peptide Y.

In addition, Peptide Y was obtained by a tryptic cleavage, and, thus, from the accepted specificity of trypsin, Peptide Y must also have lysine or arginine as its carboxy terminal amino acid. With this constraint, the allowed library of linear peptides for Peptide Y is constructed from all individual linear permutations of the combinations above. The allowed library includes over 20,000 peptides, and is thus not shown.

As with Example 1, the method of U.S. Patent No. 5,538,897 was then used to match Peptide Y to this library by tandem mass spectrometry. The experimental tandem mass spectrum of Peptide Y is shown in Fig. 8, and the top 10 ranking peptides matched to this spectrum are given below. Of these ten, the top ranking peptide, YGGFIRR is known to be Peptide Y.

Rank/Sp	(M+H)	Cn	C^4	Sp	Ions	Reference	Peptide
1/3	868.	51.000	1.894	376.6	11/24	p(415)	(-)YGGFFIR
2/1	868.5	0.967	1.831	440.4	11/24	p(298)	(-)YGGRIFR
3/15	868.5	0.966	1.830	322.8	11/28	p(1975)	(-)YGGFIGVR
4/15	868.5	0.965	1.828	322.8	11/28	p(1735)	(-)YGGFIVGR
5/5	868.5	0.961	1.821	361.7	11/24	p(454)	(-)YGGRFIR
6/2	868.5	0.960	1.819	408.0	11/24	p(1311)	(-)YGVNIFR
7/12	868.5	0.951	1.802	333.7	11/24	p(1527)	(-)YGVNFIR
8/8	868.5	0.942	1.783	356.9	11/28	p(2153)	(-)YGGGVIFR
9/13	868.5	0.937	1.775	331.0	11/24	p(394)	(-)YGGIFRR
10/8	868.5	0.935	1.771	356.9	11/28	p(2147)	(-)YGGVGIFR

EXAMPLE 3.

The amino acid sequence of Peptide Z, a known standard peptide, was determined using the method of the invention, as applied to Peptide X in Example 1 and Peptide Y in Example 2. Peptide Z has the following amino acid sequence: RPPGFSPFR. The molecular mass of Peptide Z was determined to be 1060.5660 to an experimental accuracy of 30 ppm from the mass spectrum shown in Fig. 9. The masses at 1181.6477, 1296.6933 and 1570.6774 are from internal standards, added to allow instrument calibration.

The set of amino acids that are possibly part of Peptide Z were then defined for consideration in the analysis. The defined set of amino acids with the molecular mass of each amino acid less the mass of the one water molecule lost during peptide bond formation is the same as those used in Examples 1 and 2.

As the mass of Peptide Z was measured as 1060.5660 to an experimental accuracy of 30 ppm, each allowed amino acid combination must therefore sum to a mass

equal to 1060.5660 ± 30 ppm. In addition, from the immonium ion region of the PSD trace from Fig. 9, shown in Fig. 10, it was determined that Peptide Z must also contain the following amino acids: phenylalanine with a characteristic mass at 120.20, arginine with a characteristic mass at 174.94, serine together with proline as deduced from the mass at 167.23, and glycine together with proline as deduced from the mass at 155.66.

Application of the equation in Example 1 was used to determine the allowed combinations of amino acids for Peptide Z, and demonstrates that only the following combinations of amino acids are allowed for Peptide Y:

PTIW + FRPSG	GAAAAR + FRPSG	GVVNK + FRPSG	AVVID + FRPSG	
WIW + FRPSF	GPPVF + FRPSG	ASPNK + FRPSG	SPVVD + FRPSG	
GQRR + FRPSG	GPVIM + FRPSG	GGAAIK + FRPSG	GVINN + FRPSG	
ANRR + FRPSG	APVVM + FRPSG	GGGPTK + FRPSG	AVVNN + FRPSG	
GOARR+FRPSG	AAIIE + FRPSG	GGGVVK + FRPSG	GGGVIN + FRPSG	
PPFR + FRPSG	GPTIE + FRPSG	GAAAVK + FRPSG	GAAAIN + FRPSG	
PIMR + FRPSG	GVVIE + FRPSG	GGASPK + FRPSG	GGAWN + FRPSG	
VIER + FRPSG	ASPIE + FRPSG	IQQQ + FRPSG	AAAAVN + FRPSG	
VNQR+FRPSG	APVTE+FRPSG	GAIQQ+FRPSG	SSPII+FRPSG	
CGVQR + FRPSG	AVVVE + FRPSG	AAVQQ + FRPSG	SPVTI + FRPSG	
AAAQR + FRPSG	GSPKK + FRPSG	AAINQ + FRPSG	GGGGGVI + FRPSG	
IIDR + FRPSG	IQQK +FRPSG	GVVNQ + FRPSG	GGGAAAAI + FRPSG	
INM + FRPSG	GAIQK + FRPSG	GGAAIQ + FRPSG	PPTTT + FRPSG	
GGINR + FRPSG	AAVQK + FRPSG	GGG VVQ + FRPSG	PVVTT+FRPSG	
GAVNR+FRPSG	GSPQK+FRPSG	AAAVQ+FRPSG	GGGGAVV + FRPSG	
GGGGIR + FRPSG	AAINY + FRPSG	GVIID + FRPSG	GGAAAAV + FRPSG	
GGGAVR + FRPSG	GPTNK + F RPSG	APTID + FRPSG	AAAAAAA + FRPSG	

These combinations constitute the allowed set of amino acid combinations for Peptide Z.

In addition, Peptide Z was obtained by a tryptic cleavage, and, from the accepted specificity of trypsin, Peptide Z must have lysine or arginine as its carboxy terminal amino acid. With this constraint, the allowed library of linear peptides for Peptide Z is

constructed from all individual linear permutations of the combinations above. The allowed library includes over 2,000,000 peptides, and is thus not shown.

As with Examples 1 and 2, the method of U.S. Patent No. 5,538,897 was then used to match Peptide Z to this library by tandem mass spectrometry. The experimental tandem mass spectrum of Peptide Z is shown in Fig. 11, and the top 10 ranking peptides matched to this spectrum provided below. Of these ten, the top ranking peptide, RPPGFSPFR is known to be Peptide Z.

Rank/Sp	(M + EI)	Cn	C A4	Sp	Ions	Reference	Peptide
1/1	1061.2	1.000	3.310	1163.5	19/24	p(135)	(-)RPPGFSPFR
2/2	1061.2	0.871	2.884	1126.6	19/24	p(120)	(-)RPPGFPSFR
3/5	1061.2	0.857	2.835	824.7	17/24	p(122)	(-)RPPFGPSFR
4/11	1061.2	0.849	2.811	692.8	16/24	p(164)	(-)RPPGFFPSR
5/4	1061.2	0.833	2.759	831.2	17/24	p(189)	(-)RPPGFFSPR
6/3	1061.2	0.831	2.749	872.9	17/24	p(131)	(-)RPPSFGPFR
7/6	1061.2	0.819	2.711	797.1	17/24	p(126)	(-)RPFGPPSFR
8/12	1061.2	0.806	2.668	674.0	16/24	p(100)	(-)RPPGPSFFR
9/13	1061.2	0.792	2.623	668.4	16/24	p(137)	(-)RFPPGSPFR
10/14	1061.2	0.782	2.588	656.5	16/24	p(138)	(-)RFGPPSPFR

While it is apparent that the invention disclosed herein is well calculated to fulfill the objectives stated above, it will be appreciated that numerous modifications and embodiments may be devised by those skilled in the art. Therefore, it is intended that the appended claims cover all such modifications and embodiments that fall within the true spirit and scope of the present invention.

CLAIMS

1. A method for determining the amino acid sequence of an unknown peptide, which comprises:

(a) determining a molecular mass and an experimental fragmentation spectrum for the unknown peptide;

(b) comparing the experimental fragmentation spectrum of the unknown peptide to theoretical fragmentation spectra calculated for a peptide library composed of all possible linear sequences of amino acids having a total mass that corresponds to the molecular mass of the unknown peptide; and

(c) identifying a peptide in the peptide library having a theoretical fragmentation spectrum that matches most closely the fragmentation spectrum of the unknown peptide.

2. The method of claim 1, wherein the molecular mass for the unknown peptide is determined with an accuracy of up to about 30 parts per million.

3. The method of claim 1 or claim 2, wherein the total mass of each of the possible linear sequences of amino acids is within the range of plus or minus about 30 parts per million of the molecular mass of the unknown peptide.

4. The method of any preceding claim, further comprising calculating an indication of closeness-of-fit between the experimental fragmentation spectrum of the unknown peptide and each of the theoretical fragmentation spectra calculated for the peptide library.

5. The method of claim 4, further comprising selecting peak values having an intensity greater than a predetermined threshold value when calculating the indication of closeness-of-fit.

6. The method of any preceding claim, further comprising normalizing the experimental fragmentation spectrum.

7. The method of any preceding claim, wherein the amino acids are selected from tryptophan, arginine, histidine, glutamic acid, glutamine, aspartic acid, leucine, threonine, proline, alanine, tyrosine, carbamido cysteine, phenylalanine, methionine, lysine, asparagine, isoleucine, cysteine, valine, serine, and glycine.

8. The method of any of claims 1 to 6, wherein the amino acids comprise non-natural amino acids or chemically modified forms of the naturally occurring amino acids.

9. The method of any preceding claim, wherein the unknown peptide has a molecular mass greater than about 1,400 Daltons.

10. The method of any preceding claim, wherein the molecular mass for the unknown peptide is determined using a mass spectrometer.

11. The method of claim 10, wherein the mass spectrometer is a time-of-flight mass spectrometer.

12. The method of claim 10, wherein the molecular mass and the fragmentation spectrum for the unknown peptide are determined using a tandem mass spectrometer.

13. A method according to any preceding claim, which additionally comprises the identification of one or more amino acids in the unknown peptide from its experimental fragmentation spectrum, or from its method of preparation, and using the one or more identified amino acids to constrain the library of all possible linear sequences.

14. The method of claim 12, wherein the spectrum has an immonium ion region, and the immonium region is used to identify one or more amino acids contained in the unknown peptide.

15. The method of claim 13 or claim 14, wherein the identification comprises comparing a known characteristic of amino acids with characteristics of the experimental fragmentation spectrum.

16. The method of any of claims 13 to 15, wherein said one or more amino acids is or includes the N-terminal or C-terminal amino acid.

17. A method of generating a library of amino acid sequences, wherein each sequence in the library represents a peptide having a molecular mass that corresponds to a single, predetermined molecular mass, which comprises

defining a set of combinations of allowed amino acids having a molecular weight that corresponds to the predetermined molecular mass; and

generating a library of all possible linear sequences of the amino acids in each combination of the set;

wherein the library is constrained by identification as defined in any of claims 13 to 16.

18. A method according to claim 17, additionally comprising the characteristic of any of claims 2 to 12.

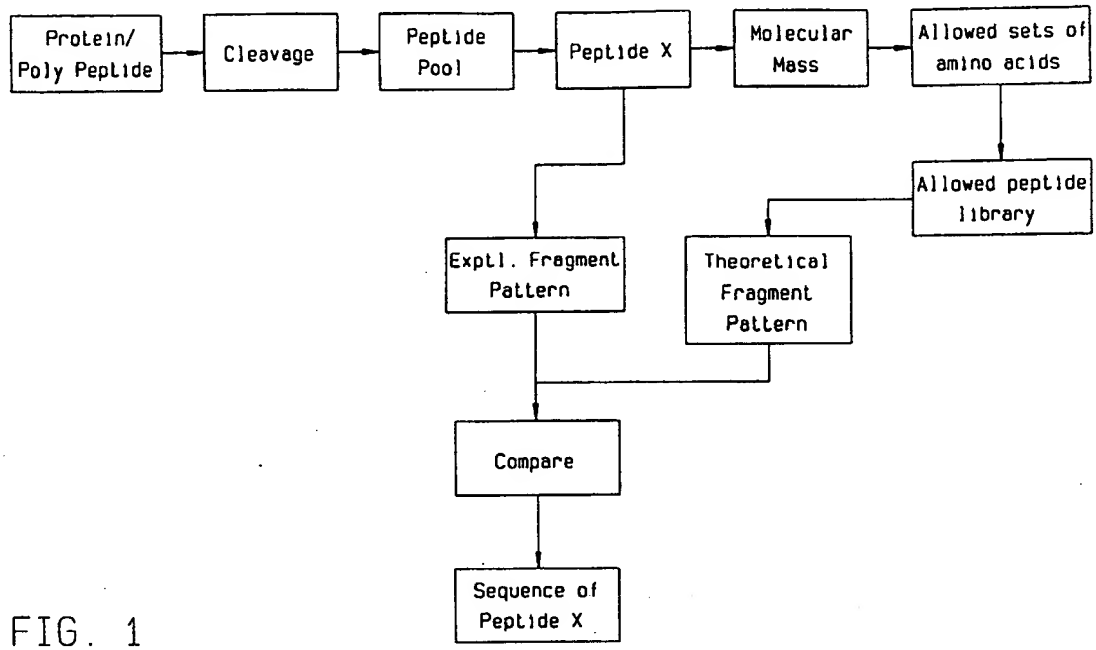

Protein/
Poly Peptide
Cleavage
Peptide
Pool
Peptide X
Molecular
Mass
Allowed sets of
amino acids
Allowed peptide
library
Exptl. Fragment
Pattern
Theoretical
Fragment
Pattern
Compare
Sequence of
Peptide X

FIG. 1

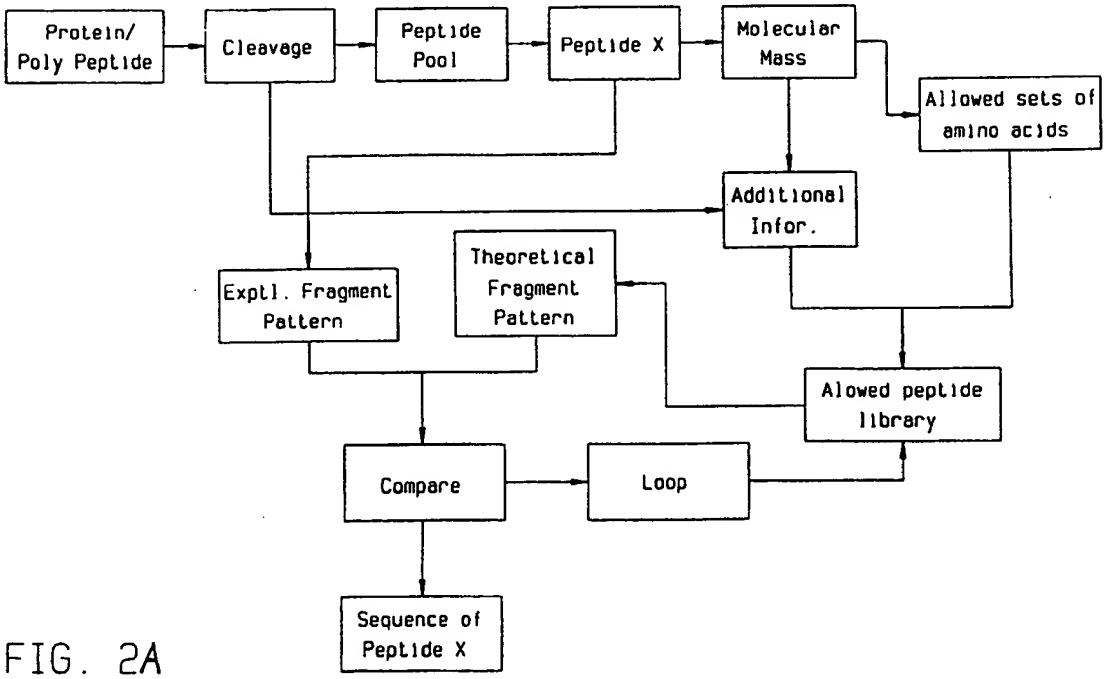

Protein/
Poly Peptide
Cleavage
Peptide
Pool
Peptide X
Molecular
Mass
Allowed sets of
amino acids
Additional
Infor.
Exptl. Fragment
Pattern
Theoretical
Fragment
Pattern
Alowed peptide
library
Compare
Loop
Sequence of
Peptide X

FIG. 2A

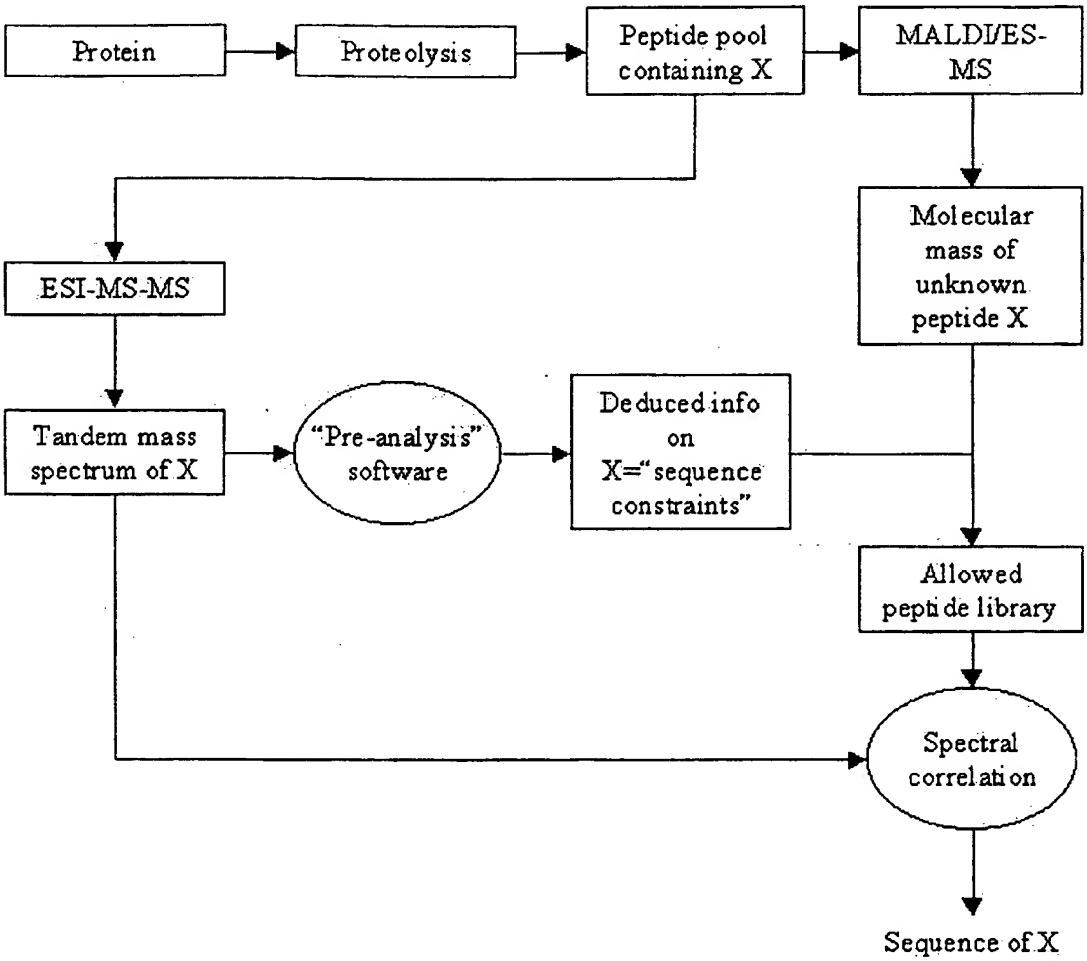

Protein
Proteolysis
Peptide pool containing X
MALDI/ES-MS
Molecular mass of unknown peptide X
ESI-MS-MS
Tandem mass spectrum of X
"Pre-analysis" software
Deduced info on X="sequence constraints"
Allowed peptide library
Spectral correlation
Sequence of X

FIG. 2B

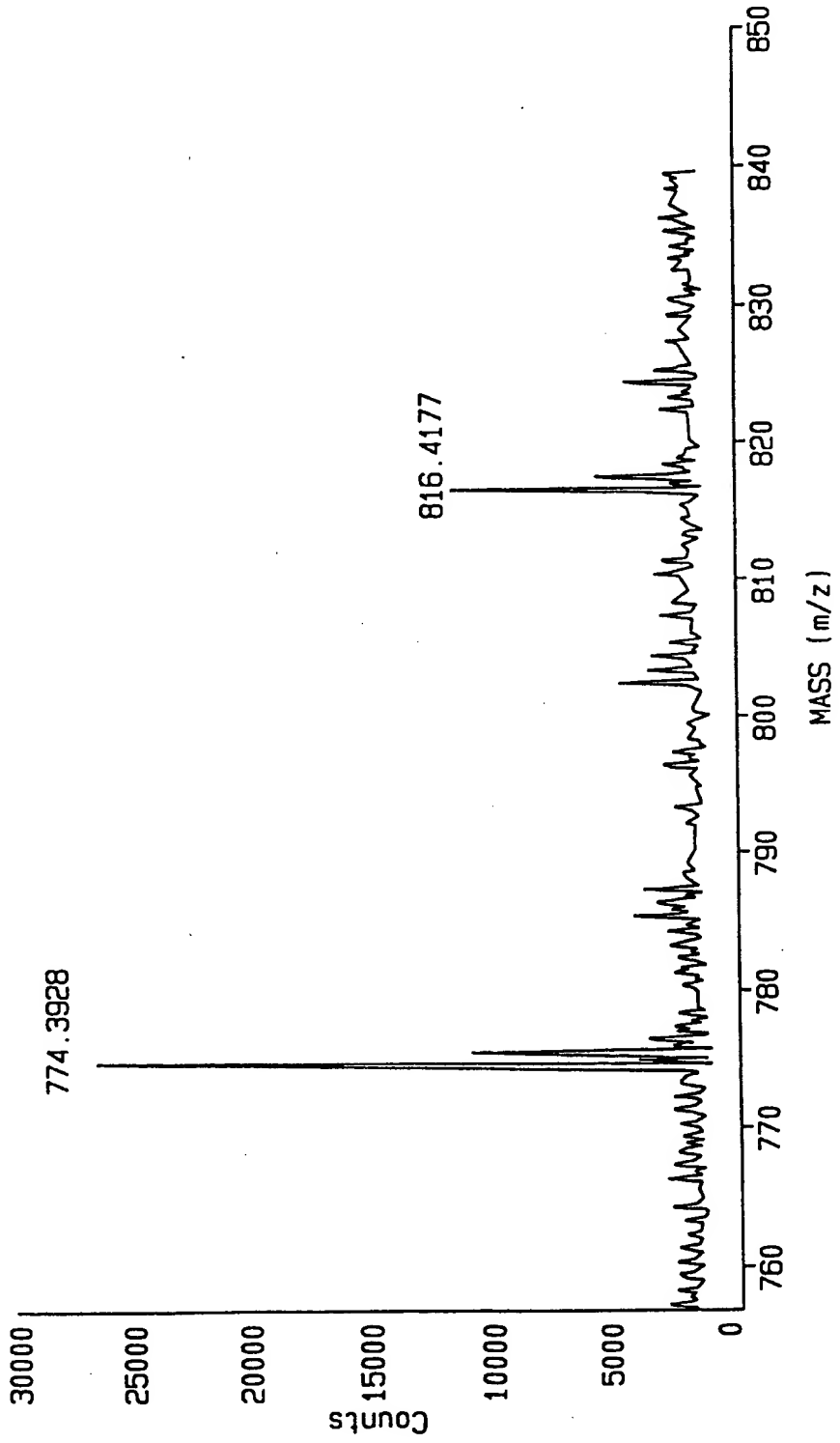

774.3928
816.4177
Counts
30000
25000
20000
15000
10000
5000
0
760
770
780
790
800
810
820
830
840
850
MASS (m/z)

FIG. 3

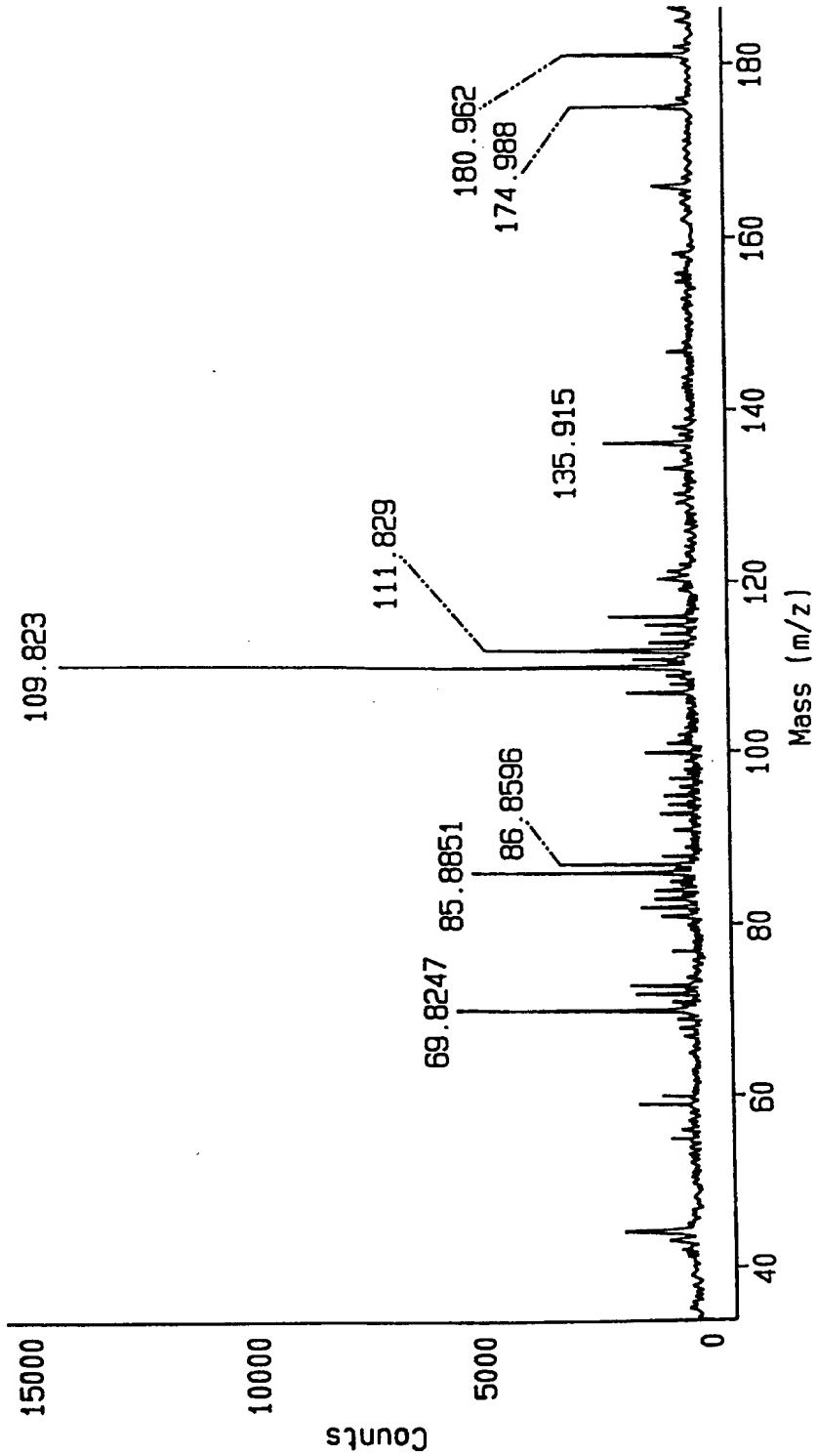

Counts
15000
10000
5000
0
40
60
80
100
120
140
160
180
Mass (m/z)
109.823
111.829
135.915
180.962
174.988
86.8596
85.8851
69.8247

FIG. 4

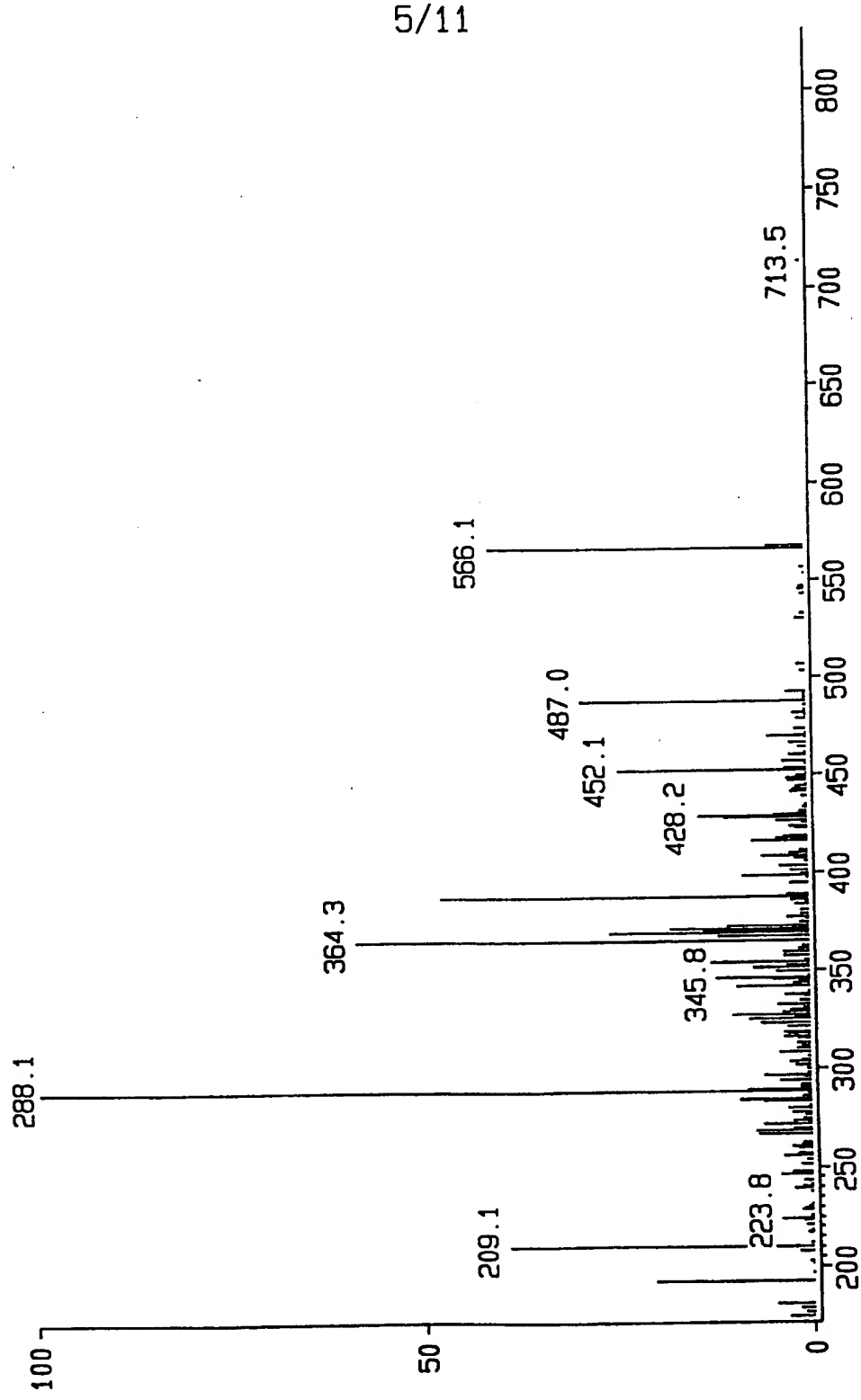

100
50
0
200
250
300
350
400
450
500
550
600
650
700
750
800
288.1
364.3
345.8
209.1
223.8
566.1
487.0
452.1
428.2
713.5

FIG. 5

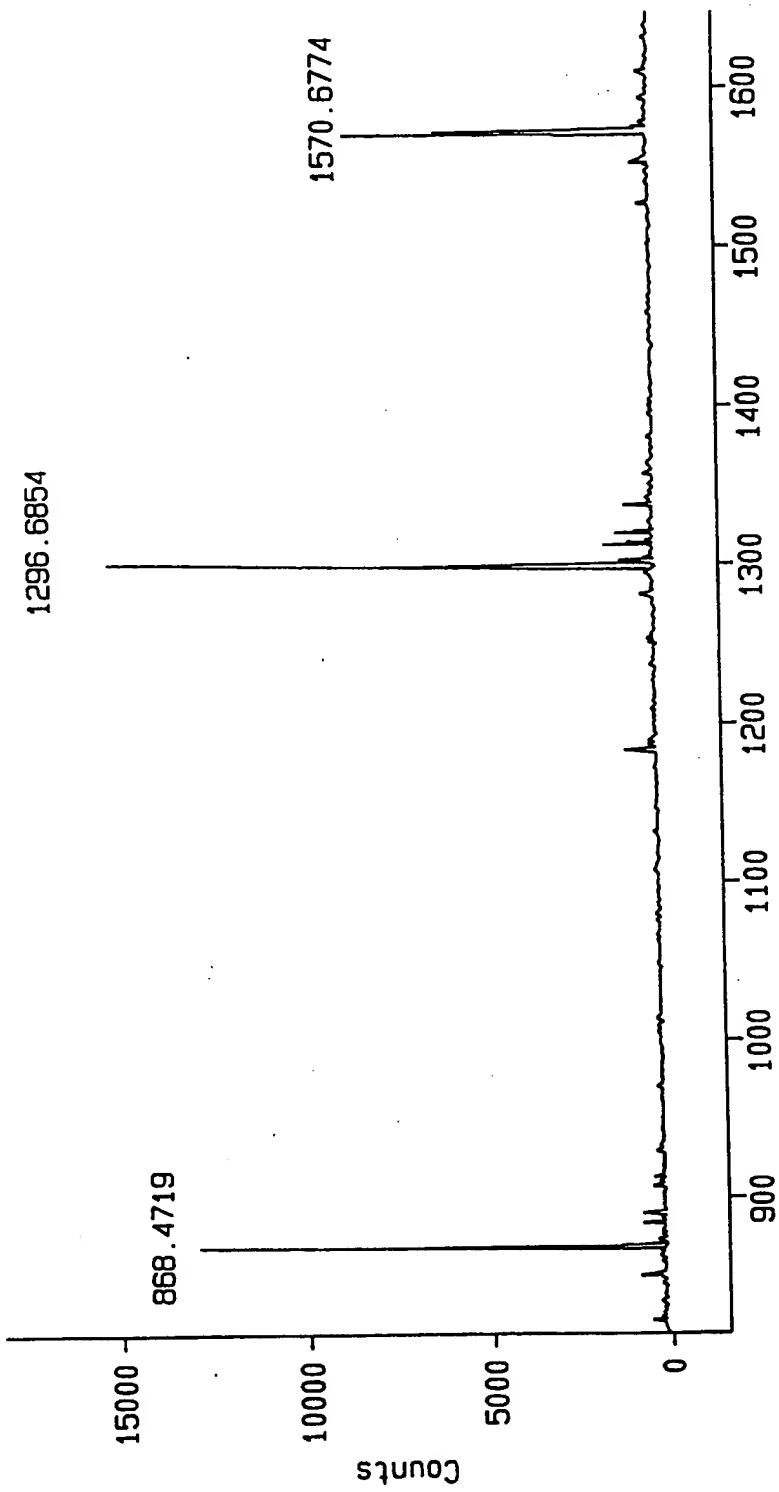

1570.6774
1296.6854
868.4719
15000
10000
5000
0
Counts
900
1000
1100
1200
1300
1400
1500
1600
Mass (m/z)

FIG. 6

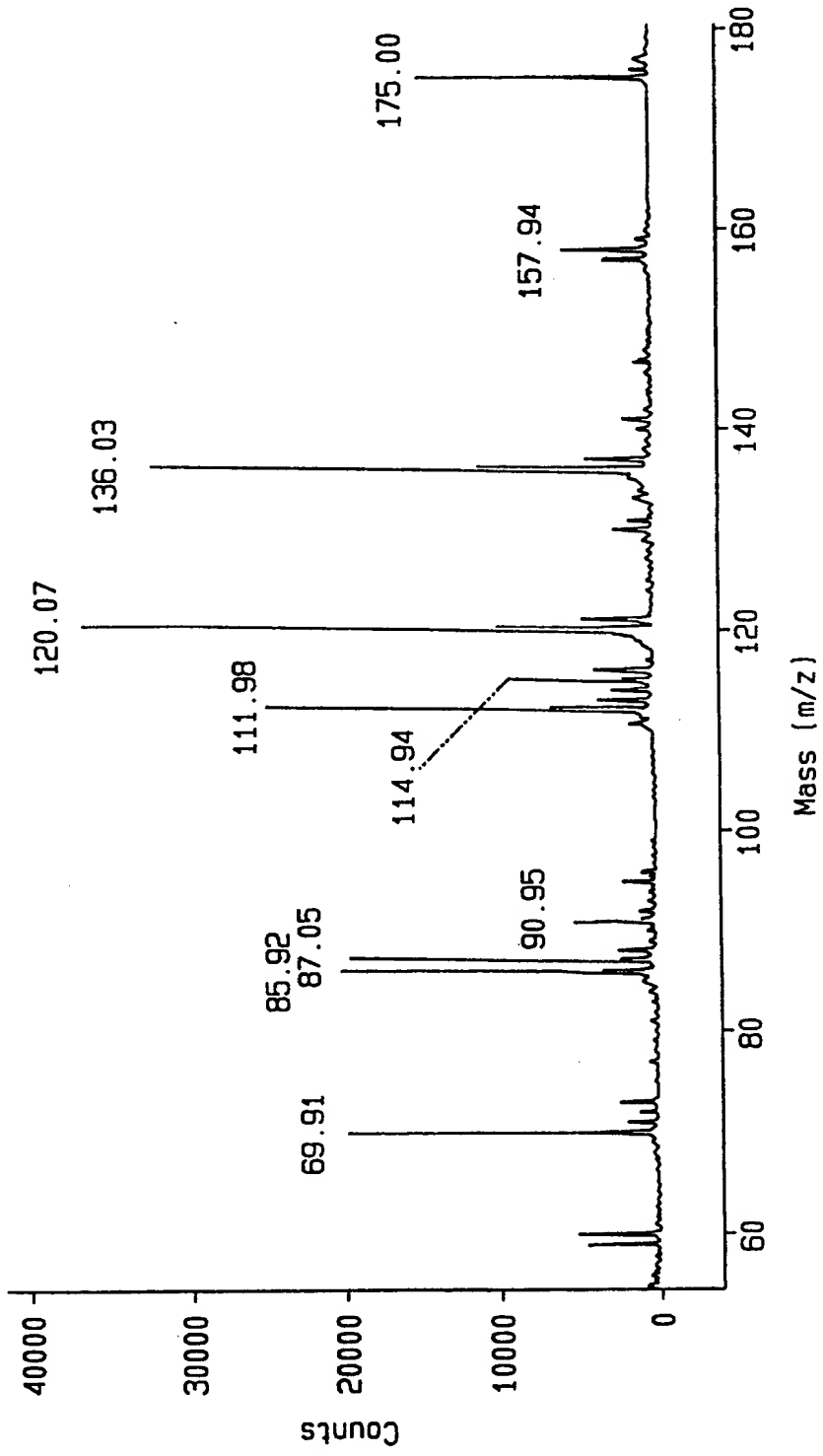

175.00
157.94
136.03
120.07
111.98
114.94
90.95
85.92
87.05
69.91
40000
30000
20000
10000
0
Counts
60
80
100
120
140
160
180
Mass (m/z)

FIG. 7

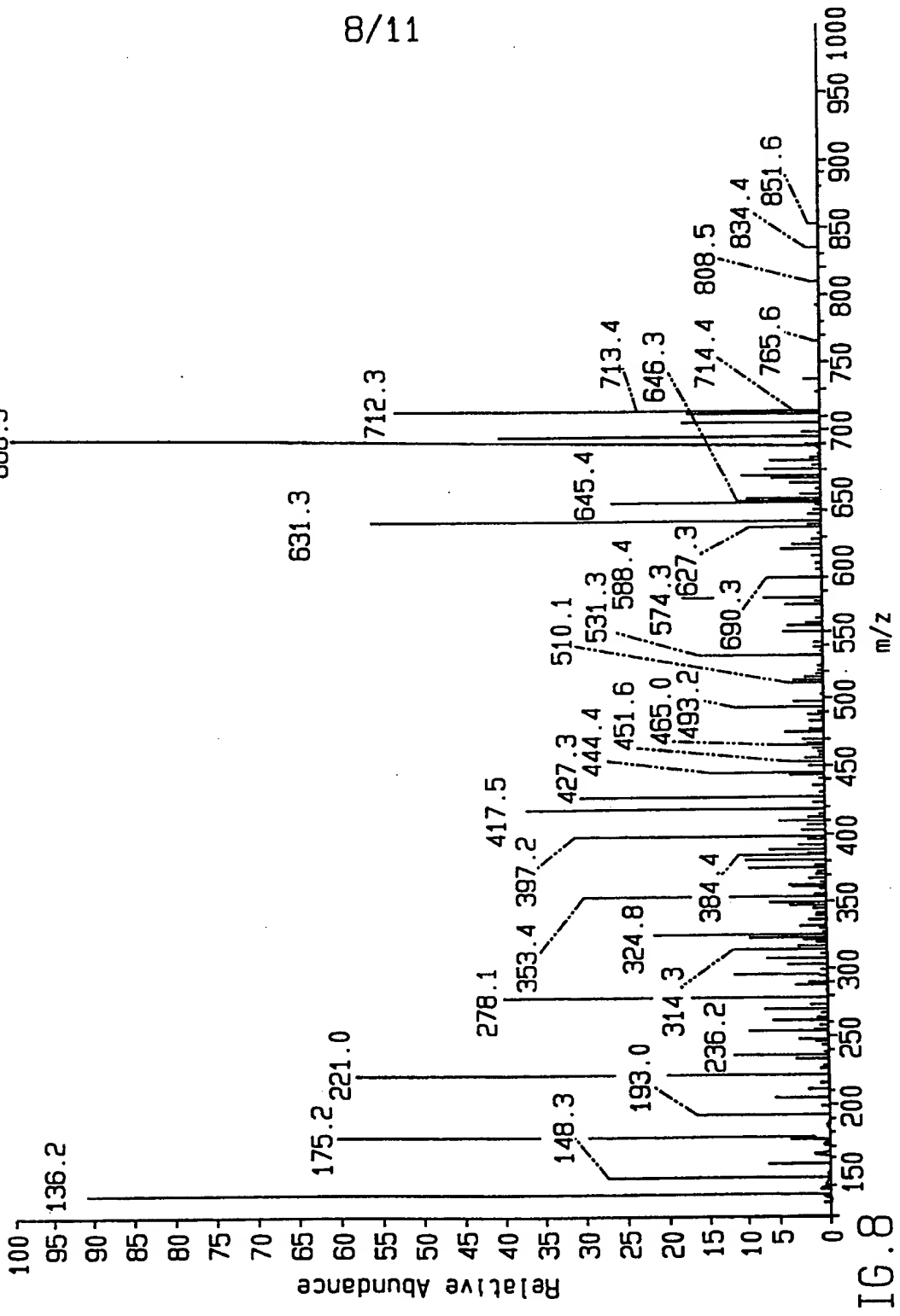

Relative Abundance
m/z
136.2
175.2
221.0
148.3
193.0
278.1
314.3
353.4
324.8
397.2
384.4
417.5
427.3
444.4
451.6
465.0
493.2
510.1
531.3
588.4
574.3
627.3
690.3
631.3
688.3
712.3
645.4
713.4
646.3
714.4
765.6
808.5
834.4
851.6

FIG.8

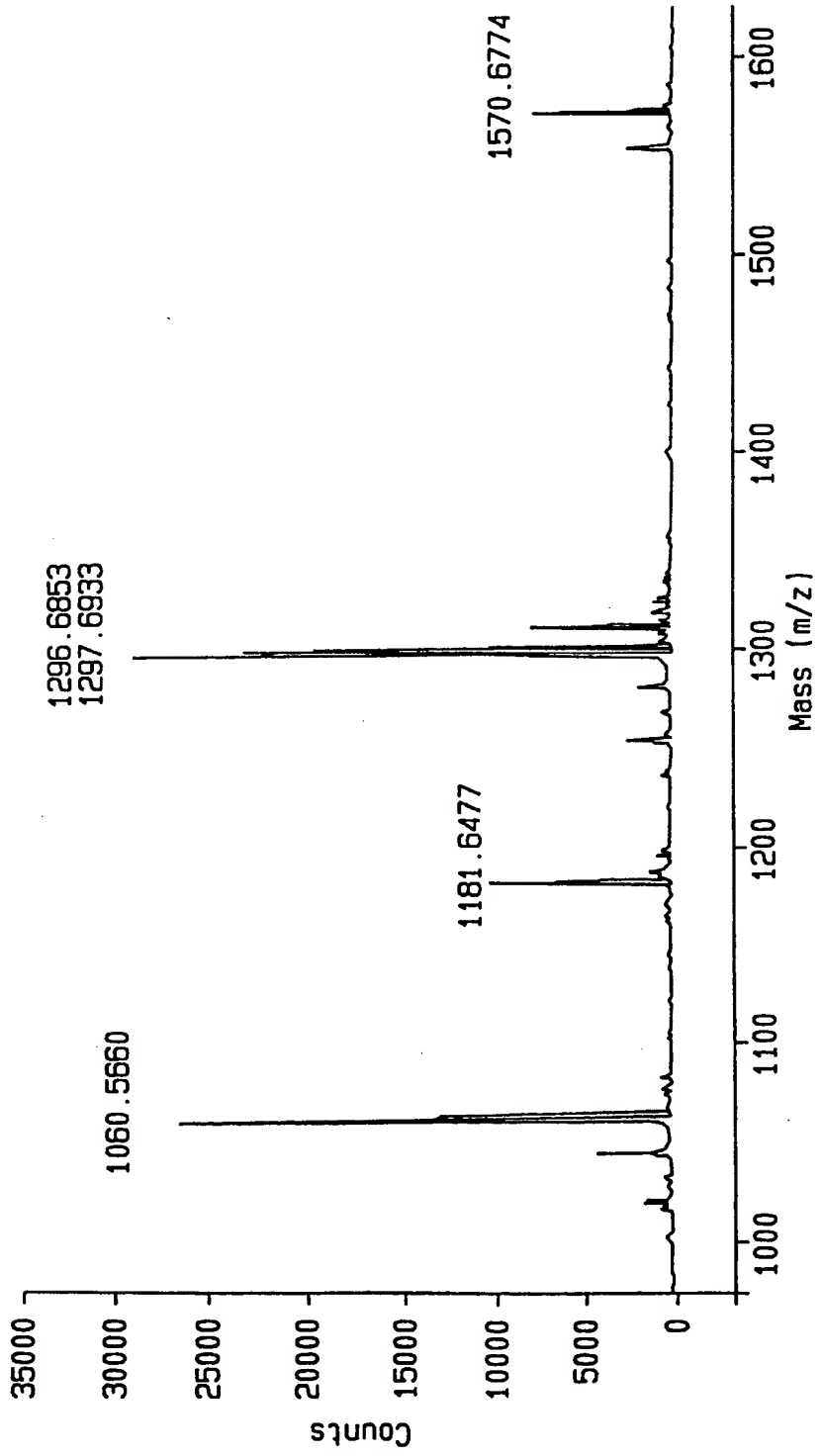

1570.6774
1296.6853
1297.6933
1181.6477
1060.5660
Counts
35000
30000
25000
20000
15000
10000
5000
0
1000
1100
1200
1300
1400
1500
1600
Mass (m/z)

FIG. 9

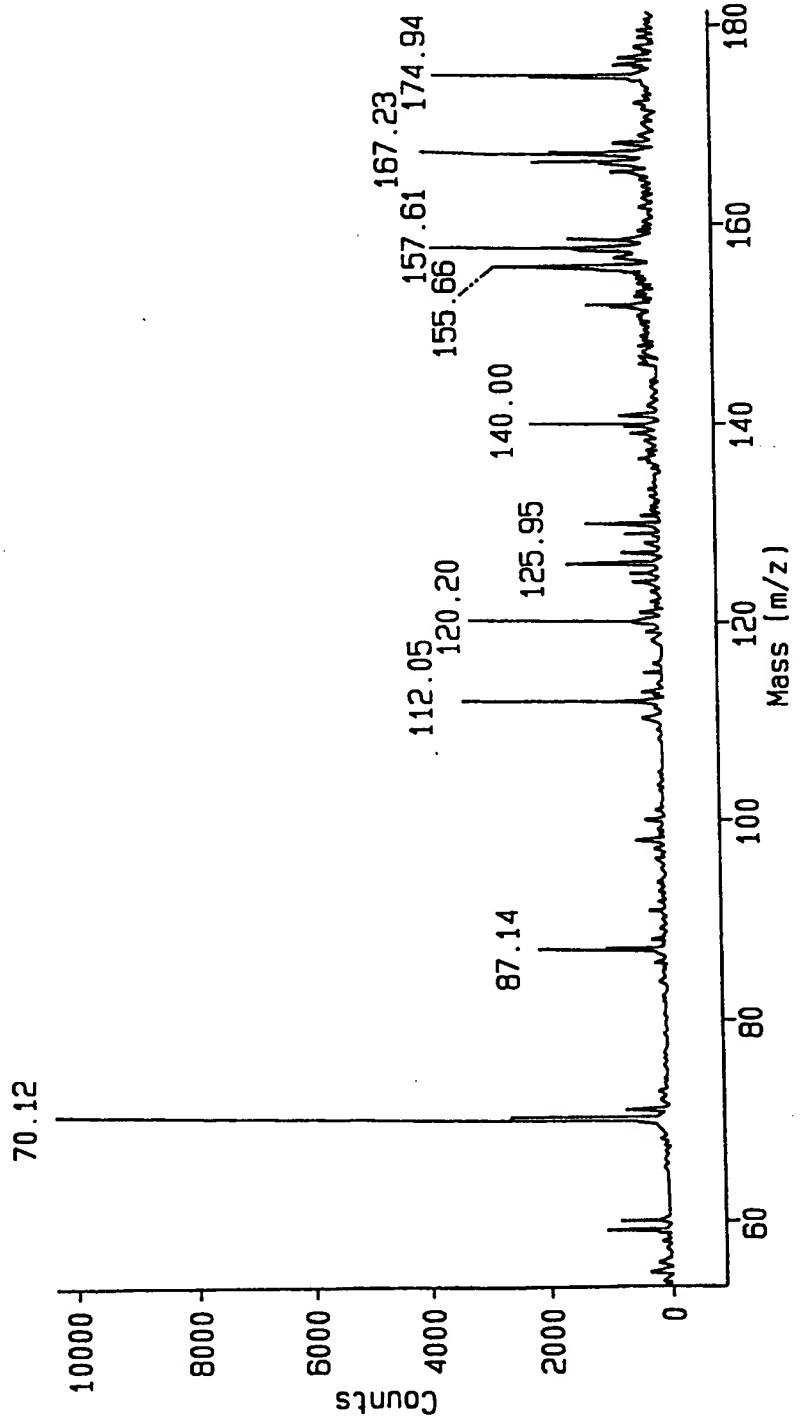

70.12
87.14
112.05
120.20
125.95
140.00
155.66
157.61
167.23
174.94
Counts
10000
8000
6000
4000
2000
0
60
80
100
120
140
160
180
Mass (m/z)

FIG. 10

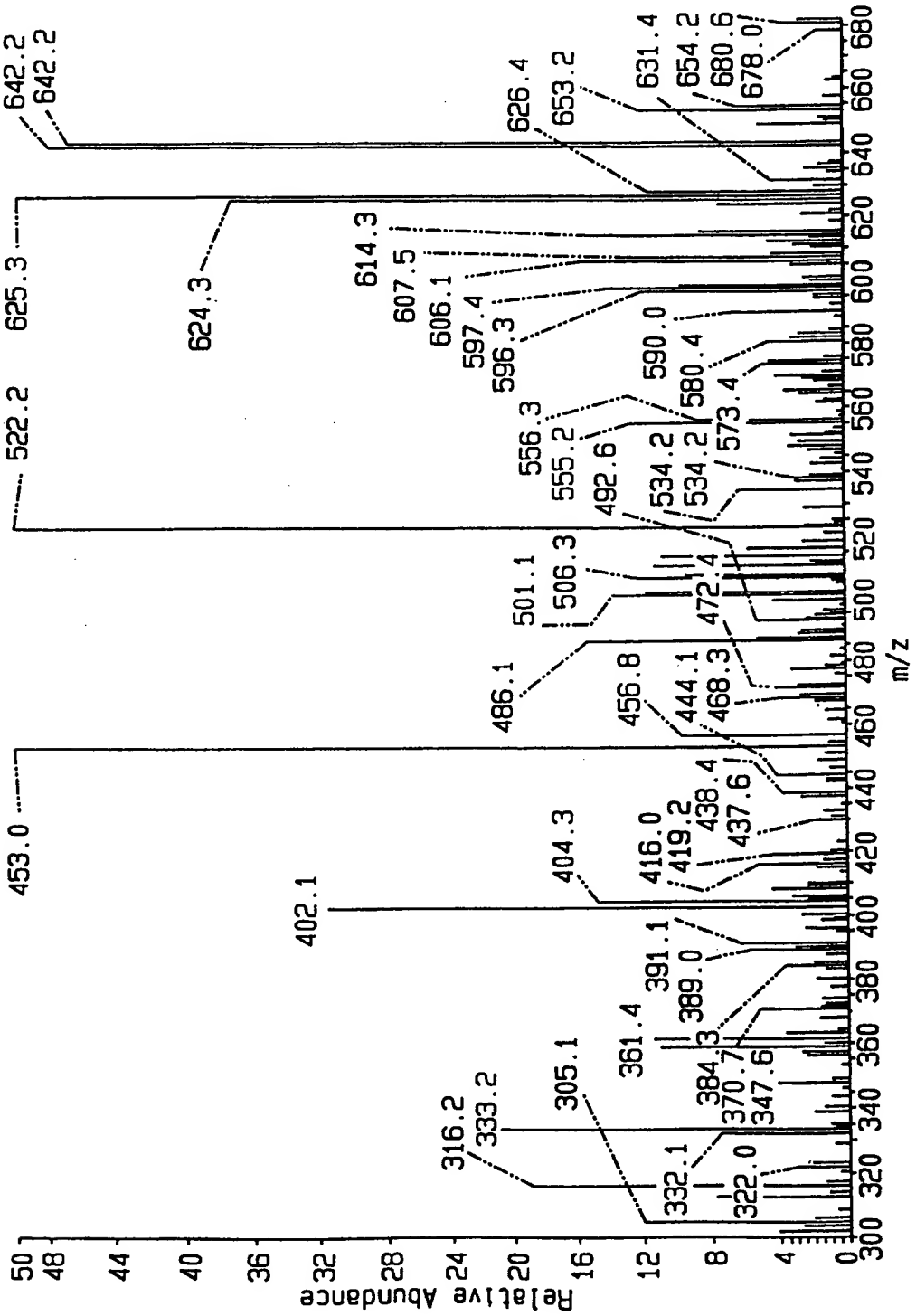

Relative Abundance
m/z
642.2
642.2
626.4
653.2
631.4
654.2
680.6
678.0
625.3
624.3
614.3
607.5
606.1
597.4
596.3
590.0
580.4
573.4
556.3
555.2
534.2
534.2
492.6
522.2
506.3
501.1
472.4
486.1
456.8
444.1
468.3
453.0
438.4
437.6
416.0
419.2
404.3
402.1
391.1
389.0
361.4
384.3
370.7
347.6
305.1
316.2
333.2
332.1
322.0

FIG. 11

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G01N 33/68 — A3

(11) International Publication Number: **WO 98/53323**

(43) International Publication Date: 26 November 1998 (26.11.98)

(21) International Application Number: PCT/GB98/01486

(22) International Filing Date: 22 May 1998 (22.05.98)

(30) Priority Data:
9710582.9 22 May 1997 (22.05.97) GB

(71) Applicant *(for all designated States except US)*: OXFORD GLYCOSCIENCES (UK) LTD. [GB/GB]; 10 The Quadrant, Barton lane, Abingdon Science Park, Abingdon OX14 3YS (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: PAREKH, Raj, Bhikhu [GB/GB]; Alchester House, Langford Lane, Nr Wendlebury, Oxfordshire OX6 0NS (GB). PRIME, Sally, Barbara [GB/GB]; Sunnybrook, 37 North Hinksey Village, Oxford OX2 0NA (GB). WEDD, Nick, Sinclair [GB/GB]; Sunnybrook, 37 North Hinksey Village, Oxford OX2 0NA (GB). TOWNSEND, Robert, Reid [US/GB]; 33 Norreys Avenue, Oxford OX1 4ST (GB).

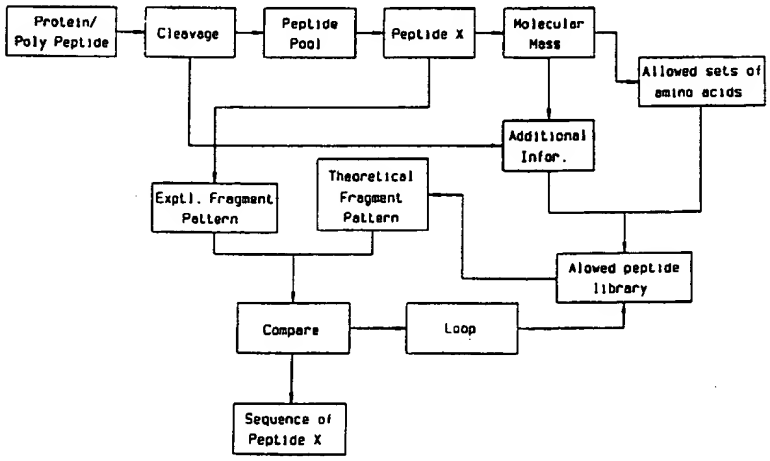
(74) Agent: GILL JENNINGS & EVERY; Broadgate House, 7 Eldon Street, London EC2M 7LH (GB).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(88) Date of publication of the international search report:
11 March 1999 (11.03.99)

(54) Title: A METHOD FOR DE NOVO PEPTIDE SEQUENCE DETERMINATION



(57) Abstract

A method for determining the amino acid sequence of an unknown peptide comprising (a) determining a molecular mass and an experimental fragmentation spectrum for the unknown peptide; (b) comparing the experimental fragmentation spectrum of the unknown peptide to theoretical fragmentation spectra calculated for a peptide library composed of all possible linear sequences of amino acids having a total mass that corresponds to the molecular mass of the unknown peptide; and (c) identifying a peptide in the peptide library having a theoretical fragmentation spectrum that matches most closely the fragmentation spectrum of the unknown peptide.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INTERNATIONAL SEARCH REPORT

I. ıational Application No
PCT/GB 98/01486

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G01N33/68

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 470 753 A (SEPETOV NIKOLAI ET AL) 28 November 1995 see column 1, line 53 - column 2, line 32 see column 3, line 32 - line 39 ---	1-16
X	US 5 538 897 A (YATES III JOHN R ET AL) 23 July 1996 cited in the application see the whole document --- -/--	1-16

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of theinternational search	Date of mailing of the international search report
26 October 1998	20/11/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Hoekstra, S

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 98/01486

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MEDZIHRADSZKY K F ET AL: "Peptide Sequence Determination by Matrix-Assisted Laser Desorption Ionization Employing a Tandem Double Focusing Magnetic-Orthogonal Acceleration Time-of-Flight Mass Spectrometer" JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY, vol. 7, no. 1, January 1996, page 1-10 XP004051936 see page 7 - page 8 ---	1-16
A	MEDZIHRADSZKY K F ET AL: "Protein sequence and structural studies employing matrix-assisted laser desorption ionization-high energy collision-induced dissociation" INTERNATIONAL JOURNAL OF MASS SPECTROMETRY AND ION PROCESSES, vol. 160, no. 1, January 1997, page 357-369 XP004058842 see the whole document ---	1-16
A	MANN M ET AL: "ERROR-TOLERANT IDENTIFICATION OF PEPTIDES IN SEQUENCE DATABASES BY PEPTIDE SEQUENCE TAGS" ANALYTICAL CHEMISTRY, vol. 66, no. 24, 15 December 1994, pages 4390-4399, XP000573399 see the whole document ---	1-16
A	WO 95 25737 A (PENN STATE RES FOUND ;BENKOVIC STEPHEN J (US); WINOGRAD NICHOLAS ( ) 28 September 1995 see the whole document ---	11
A	J A BOUTIN, P HENNIG, P-H LAMBERT, S BERTIN, L PETIT, J-P MAHIEU, B SERKIZ, J-P VOLLAND, J-L FAUCHÈRE: "Combinatorial Peptide Libraries: Robotic Synthesis and Analysis by Nuclear Magnetic Resonance, Mass Spectrometry, Tandem Mass Spectrometry, and High-Performance Capillary Electrophoresis Techniques" ANALYTICAL BIOCHEMISTRY, vol. 234, 1996, pages 126-141, XP002081905 see the whole document --- -/--	1-16

2

INTERNATIONAL SEARCH REPORT

Ir ational Application No
PCT/GB 98/01486

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	R S YOUNGQUIST, G R FUENTES, M P LACEY, T KEOUGH: "Generation and Screening of Combinatorial Peptide Liraries Designed for Rapid Sequencing by Mass Spectrometry" JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, vol. 117, no. 14, 1995, pages 3900-3906, XP002081906 see the whole document ---	1-16
A	M A KELLY, H LIANG, I-I SYTWU, I VLATTAS, N L LYONS, B R BOWEN, L P WENNOGLE: "Characterization of SH2-Ligand Interactions via Library Affinity Selection with Mass Spectrometric Detection" BIOCHEMISTRY, vol. 35, no. 36, 1996, pages 17747-11755, XP002081907 see the whole document -----	1-16

2

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB 98/01486

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. [X] Claims Nos.: 17,18
because they relate to subject matter not required to be searched by this Authority, namely:
Rule 39.1(v) PCT-Presentation of information
Rule 39.1(iii) PCT-
Scheme,rules and method for performing mental acts
Rule 39.1(vi) PCT -

2. [ ] Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. [ ] Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. [ ] As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. [ ] As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. [ ] As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. [ ] No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

[ ] The additional search fees were accompanied by the applicant's protest.

[ ] No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

In ational Application No
PCT/GB 98/01486

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 5470753	A	28-11-1995	AU	4844893 A	29-03-1994
			WO	9406017 A	17-03-1994
US 5538897	A	23-07-1996	CA	2185574 A	21-09-1995
			EP	0750747 A	02-01-1997
			JP	9510780 T	28-10-1997
			WO	9525281 A	21-09-1995
WO 9525737	A	28-09-1995	EP	0751950 A	08-01-1997
			JP	9510711 T	28-10-1997